Prozessorientiertes Assessment im Bereich der Lesekompetenzdiagnostik

Eine metaanalytische Untersuchung und Verfahrenskonstruktion

Von der Pädagogischen Hochschule Heidelberg
zur Erlangung des Grades einer
Doktorin der Philosophie (Dr. phil.)
genehmigte Dissertation von

Christine M. Marx, geb. Chilla
aus
Wertingen

2016

Erstgutachter:	Prof. Dr. Tobias Dörfler (Pädagogische Hochschule Heidelberg)
Zweitgutachter:	Prof. Dr. Markus Vogel (Pädagogische Hochschule Heidelberg)
Fach:	Psychologie
Tag der Mündlichen Prüfung:	14.12.2016

Hinweise:

Diese Dissertation erfolgte im Rahmen des Projekts „*Prozessorientierte Diagnostik der Lesekompetenz in der Grundschule*“. Das Projekt wurde aus Mitteln der Pädagogischen Hochschule Heidelberg, Kostenstelle 101030701, gefördert. Die empirischen Datenerhebungen an den teilnehmenden Schulen wurden vom Regierungspräsidium Karlsruhe, den Schulleitern der teilnehmenden Schulen und den Erziehungsberechtigten der teilnehmenden Kinder genehmigt. Die Teilnahme war anonym und erfolgte auf freiwilliger Basis.

In dieser Arbeit wird durchgehend lediglich die männliche Form verwendet, wenn sowohl die männliche als auch die weibliche Form gemeint sind. Dies geschieht ausschließlich aus Gründen der Lesbarkeit. Eine Wertung soll damit in keiner Form impliziert sein.

Für

Pia Chilla

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ACC	*Accuracy*-Daten
FR	Feedbackresponsivität
IRT	*Item-Response*-Theorie
KTT	Klassische Testtheorie
RT	Reaktionszeiten
ZDP	Zone der nächsten Entwicklung

Abkürzungen im Zusammenhang mit der Testitembezeichnung

B	Item(s) mit Brückeninformationen
E	Item(s) mit expliziter Aufgabenart
I	Item(s) mit impliziter Aufgabenart
K	Item(s) mit kausaler Informationsart
L	Item(s) mit lokaler Informationsart
N	Item(s) mit einem narrativen Text als Aufgabenstamm
P	Item(s) mit paraphrasierter Aufgabenart
S	Item(s) mit einem Sachtext als Aufgabenstamm
T	Item(s) mit temporaler Informationsart

Abkürzungen in der metaanalytischen Untersuchung

DA	Dynamisches Assessment gesamt (DA_A und DA_B)
DA_A	Allgemeines, auf kognitive Fähigkeiten abzielendes dynamisches Assessment
DA_B	Bereichsspezifisches, auf Lesen abzielendes dynamisches Assessment
LK	Lesekompetenz gesamt (LK_T und LK_{NT})
LK_{NT}	Nicht hierarchiehohe, nicht auf die Ebene des Textes abzielende Lesekompetenzkomponenten
LK_T	Hierarchiehohe, auf die Ebene des Textes abzielende Lesekompetenzkomponente

1 Ziel und Struktur der vorliegenden Arbeit

Die vorliegende Arbeit ist als ein Brückenschlag zu verstehen, bei dem dynamische Prinzipien im Bereich der Leseforschung Anwendung finden sollen. Ein Ziel der vorliegenden Arbeit ist die Konstruktion und Validierung eines dynamischen Lesekompetenztests für Kinder der dritten und vierten Jahrgangsstufe, wobei Kinder mit spezifischem Förderbedarf besonders berücksichtigt werden sollen.

Zwei Konstrukte spannen den theoretischen Rahmen dieser Arbeit auf: der Begriff des dynamischen Assessments (*Kapitel 2*) und der Begriff der Lesekompetenz (*Kapitel 3*). Hierbei liegt der Schwerpunkt der theoretischen Ausführungen auf der Beschreibung der Lesekompetenz, da sie die Grundlage für die spätere Testmaterialentwicklung darstellt. Von besonderer Bedeutung ist in diesem Zusammenhang auch die Abgrenzung der Konstrukte untereinander, sie soll daher auf ein empirisches Fundament gestellt werden. Eine systematische Übersicht über ausgewählte empirische Arbeiten, welche die Bereiche des dynamischen Assessments und der Lesekompetenz miteinander verbinden, findet sich in *Kapitel 4*. In ihm wird in einer metaanalytischen Untersuchung der Zusammenhang der beiden Konstrukte, die im zu entwickelnden Lesekompetenztest erhoben werden sollen, empirisch eruiert. An die so gewonnenen Erkenntnisse schließt ein eigener Versuch an, ein Messinstrument zu entwickeln, welches eine valide dynamische Erfassung der Lesekompetenz ermöglichen soll.

Das Vorgehen der Testkonstruktion orientiert sich an den Empfehlungen von Pospeschill (2010) zur Entwicklung eines psychologischen Tests. Ausgehend von den theoretischen Überlegungen und dem momentanen Forschungsstand können die zu erfassenden Konstrukte genauer präzisiert werden, wobei hier eine zielführende Ein- und Abgrenzung dieser Konstrukte anvisiert wird.

Die Spezifizierung der Konstrukte, die im zu konstruierenden Test gemessen werden sollen, stellen die Basis für die Charakteristiken dar, nach denen das

Testmaterial systematisch entworfen und entwickelt wird (*Kapitel 5*). Das so konstruierte Testmaterial soll zunächst explorativ-qualitativ an der Zielpopulation erprobt und gegebenenfalls verbessert werden (*Kapitel 6*). Anschließend wird die Pilotierung durchgeführt, bei der die Testversion an Hand einer großen Stichprobe erstmals quantitativ überprüft werden soll (*Kapitel 7*). Auf Basis der Pilotierung kann eine vorläufige Testendversion erstellt werden, die an Hand von externen Außenkriterien validiert werden soll (*Kapitel 8*). Durch dieses standardisierte Vorgehen soll eine systematische Genese des Tests sichergestellt werden. Diese trägt wesentlich zu den psychometrisch hochwertigen Eigenschafen des Instruments bei, die eine fundierte psychologische Diagnostik auf metrisch-vergleichender Ebene erst möglich machen (Pospeschill, 2010).

Eine zusammenfassende Würdigung (*Kapitel 9*) rundet die Arbeit ab.

2 Dynamisches Assessment

2.1 Definition und Komponenten

Dynamische Testdiagnostik kann verstanden werden als ein „von Guthke und Wiedel [*sic*] (1996) eingeführter Sammelbegriff für alle testdiagnostischen Strategien, die über die gezielte Evozierung und Erfassung der intraindividuellen Variabilität im Testprozess entweder auf eine validere Erfassung des aktuellen und tatsächlichen Standes eines psychischen Merkmals und/oder seiner Veränderbarkeit abzielen“ (Häcker, 2013). Sie ist somit als heterogene Begrifflichkeit aufzufassen, die verschiedene Ausrichtungen und Ausführungen umfasst. Die einzelnen dynamischen Verfahren, ihre Klassifikation und Unterscheidungsmerkmale werden ausführlich in den Kapiteln 2.4 und 2.5 thematisiert. Nachfolgend sollen in dieser Arbeit die Begriffe dynamischer Test, dynamische Testung und dynamisches Assessment sowie die Begriffe Lernfähigkeit und Lernpotential als bedeutungsgleiche Begrifflichkeiten verwendet werden.

Zentral für alle Arten des dynamisches Assessments ist, dass zwei unterschiedliche Komponenten erfasst werden (Beckmann, 2001; Guthke & Wiedl, 1996): die Performanz/Merkmalsausprägung in bestimmten Bereichen und die Veränderung/Veränderbarkeit des Merkmals, die die dynamische Komponente darstellt und nachfolgend auch als Lernfähigkeit bezeichnet werden soll. Die Erfassung der Lernfähigkeit ist das Alleinstellungsmerkmal des dynamischen Assessments und unterscheidet es von herkömmlichen, statischen Testverfahren. Sie impliziert, dass das zu erfassende Merkmal generell modifizierbar ist (Sternberg, 2004). Daraus folgt, dass dynamische Tests nicht bei Merkmalen umgesetzt werden können, welche als nicht veränderbar charakterisiert sind. Es kann davon ausgegangen werden, dass die Auswirkungen des dynamischen Assessments auf das zu messsende Merkmal nicht geschlechtsspezifisch sind (Ramazanpour, Nourdad & Nouri, 2016).

Die Dichotomie der Merkmalserfassung findet auch Eingang in die Konstruktionsprinzipien dynamischer Tests. So hat sich bei der Konstruktion

dynamischer Verfahren unter anderem auch der Ansatz bewährt, ursprünglich statische Testverfahren um eine dynamische Testkomponente zu erweitern (z. B. Jacobs, 2001; Guthke & Gitter, 1991).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Dynamische Verfahren zielen neben der Erfassung eines Merkmals auch auf seine Veränderbarkeit ab, damit muss die Messung zweier Konstrukte im zu konstruierenden dynamischen Lesekompetenztest umgesetzt werden. Neben der zum Testzeitpunkt aktuellen Merkmalsausprägung soll auch die Veränderbarkeit des Merkmals im zu entwickelnden Test als Konstrukt berücksichtigt werden.

Bei der Konstruktion dynamischer Verfahren kann eine zunächst statische Testversion gezielt um eine dynamische Komponente erweitert werden. Eine Voraussetzung für eine gelingende Konstruktion einer dynamischen Testerweiterung ist eine prinzipielle Veränderbarkeit des zu erfassenden Merkmals. Eine weitere Voraussetzung ist ein Verständnis für das Konzept der Lernfähigkeit, das eine Abgrenzung zu ähnlichen Konzepten und eine genaue Spezifizierung zulässt. Daher soll die Lernfähigkeit zunächst in Kapitel 2.2 in einen größeren theoretischen Rahmen eingebettet werden. Auch ist zu überprüfen, ob und unter welchen Umständen ein dynamisches Verfahren in der empirischen Praxis sinnvoll umgesetzt werden kann (Kapitel 2.3). Die gelingende Konstruktion eines dynamischen Testverfahrens kann insbesondere dann angenommen werden, wenn das dynamische Assessment als in der Empirie bewährt angesehen werden kann.

2.2 Theoretische Einbettung

Die Idee der dynamischen Testverfahren hat sich unabhängig voneinander in verschiedenen Ländern entwickelt, so dass verschiedene Wissenschaftler als „Vater der dynamischen Testung“ betrachtet werden können, beispielsweise Alfred Binet (USA), Reuven Feuerstein (Israel) und Lev Wygotski (UdSSR) (Murphy, 2011, S.3). Für das Konzept des dynamischen Assessments in der vorliegenden Arbeit hat Wygotskis Ansatz besondere Relevanz.

Ansatz von Wygotski

Der Beitrag Wygotskis zum dynamischen Testen ist eingebettet in einen größeren theoretischen Rahmen, dem soziokulturellen Ansatz. Nach diesem Ansatz sind psychische Prozesse zunächst soziale und äußere Prozesse, bevor sie verinnerlicht und damit erst zu psychischen Prozessen werden. Sie haben somit einen sozialen Ursprung (Brandes, 2013). Für eine genauere Darstellung der soziokulturellen Entwicklung nach Wygotski sei beispielsweise auf Poehner (2008) verwiesen.

Als ein für die vorliegende Arbeit zentrales Konzept im Ansatz von Wygotski kann die „Zone der nächsten Entwicklung“ (*zone of proximal development*, ZDP) angesehen werden (Wygotsky, 1978). Die ZDP wird von Sternberg und Grigorenko (2002) wie folgt zusammengefasst: „the ZPD reflects development itself: It is not what one is, but what one can become; it is not what has developed, but what is developing“ (Sternberg & Grigorenko, 2002, S. 37). Damit ist die ZPD ein Konstrukt, das großen Spielraum für Interpretationen zulässt. Sie ist bislang empirisch noch nicht gut abgesichert (Sternberg & Grigorenko, 2002, S. 39). Durch die Interaktion zwischen dem Kind und seinem sozialen Umfeld (Erwachsene oder fortgeschrittenere Gleichaltrige) kann ausgehend vom aktuellen Entwicklungsstand die Zone der nächsten Entwicklung erreicht werden. Sie ist damit ein Maß für das Lernpotential relativ zum momentanen Entwicklungsstand (Rapp, 2013). Die induzierte Veränderung ist nach diesem Ansatz immer positiv. Eine mögliche negative Auswirkung der Intervention auf den Lerner wird nicht berücksichtigt (van der Veer & Valsiner, 1994, S. 6).

Die Zone der nächsten Entwicklung weist somit Parallelen zum Stufenmodell von Piaget auf, im Gegensatz zum Ansatz von Piaget läuft die Hauptrichtung der Entwicklung jedoch vom Sozialen zum Individuellen, nicht vom Individuellen zum Sozialisierten (Brandes, 2013). Wygotskis Ansatz gibt somit dem sozialen Umfeld des Kindes ein großes Gewicht, diese starke soziale Determinierung entsprach dem Zeitgeist der Sowjetunion der damaligen Zeit (Grigorenko, 2004). Insofern sind die Ansätze von Wygotski und Piaget nicht in direkter Konkurrenz zueinander zu sehen. Vielmehr ergänzen sie einander (Bliss, 1996).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Das Prinzip der dynamischen Testung kann als in der psychologischen Theorie hinreichend verankert angesehen werden, was nach Pospeschill (2010) eine besonders gute theoretische Voraussetzung für die Testkonstruktion darstellt. Damit ist das dynamische Assessment aus theoretischer Sicht hinreichend legitimiert. Für die gelingende Konstruktion eines dynamischen Verfahrens ist es jedoch darüber hinaus noch von Relevanz, ob sich dynamische Testverfahren als empirischer Forschungsgegenstand generell bewährt haben.

2.3 Dynamisches Assessment in ausgewählten empirischen Untersuchungen

In einer Vielzahl empirischer Studien werden dynamische Testverfahren angewandt. So wird das dynamische Assessment bislang in verschiedensten Bereichen eingesetzt, wie beispielsweise im Bereich der Sprache und des Lesens (Mehri & Amerian, 2015; Mardani & Tavakoli, 2011; Pishghadam, Barabadi & Kamrood, 2011; Swanson, 2010; Swanson & Howard, 2005; Pena, Iglesias & Lidz, 2001). Insgesamt hat es sich im Rahmen empirischer Studien bewährt. So zeigen in einer Übersichtsarbeit von Murphy und Maree (2006) 21 von 29 berücksichtigten Studien, dass die Leistung eines Lerners durch das dynamische Assessment mindestens so gut vorhergesagt werden kann wie durch konventionelle statische Testverfahren. Lediglich zwei Studien kommen zu dem Ergebnis, dass das dynamische Assessment den statischen Testverfahren in dieser Hinsicht unterlegen ist.

Empirische Befunde zeigen außerdem, dass dynamische Verfahren bei Kindern und insbesondere auch im Primarstufenbereich erfolgreich implementiert werden können (z. B. Gustafson, Svensson & Fälth, 2014; Resing, Stevenson & Bosma, 2012; Resing, Xenidou-Dervou, Steijn & Elliott, 2012; Fuchs, Compton, Fuchs, Bouton & Caffrey, 2011; Jeltova et al., 2011; Resing, Tunteler, de Jong & Bosma, 2009; Grigorenko et al., 2006; Sternberg et al., 2002; Klauer & Sydow, 1992).

Vielfach bewährt hat sich dabei der Einsatz dynamischer Testverfahren bei unterdurchschnittlicher Leistung und bei Schülern mit Förderbedarf (z. B. Gustafson et al., 2014; Lawrence & Cahill, 2014; Resing, Stevenson et al., 2012; Swanson, 2011; Swanson, 2010; Peltenburg, van den Heuvel-Panhuizen & Doig, 2009; Swanson & Howard, 2005; Rutland & Campbell, 1995; Swanson, 1995; Speece, Cooper & Kibler, 1990; Samuels, Tzuriel & Malloy-Miller, 1989; Tzuriel & Klein, 1985; Wiedl & Carlson, 1981).

Insbesondere hat sich hierbei gezeigt, dass das dynamisches Assessment computeradministriert umgesetzt werden kann (z. B. Golke, Dörfler & Artelt,

2015; Darhower, 2014; Pishghadam et al., 2011; Beckmann, Beckmann & Elliott, 2009; Pea, 2004; Elliott, 2003; Tzuriel & Shamir, 2002).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Empirisch hat sich das dynamische Assessment vielfach bewährt, was für die gelingende Konstruktion eines dynamischen Lesekompetenztests spricht. Es kann außerdem davon ausgegangen werden, dass ein dynamischer Test für Schüler der Primarstufe unter besonderer Berücksichtigung der Kinder mit spezifischem Förderbedarf erfolgreich umgesetzt werden kann. Die computergestützte Umsetzung einer dynamischen Testung erscheint ebenfalls vielversprechend.

Das dynamische Assessment kann somit als durch ein theoretisches Fundament und eine breite empirische Basis hinreichend legitimiert angesehen werden, um eine erfolgsversprechende Verfahrenskonstruktion wahrscheinlich zu machen. Um jedoch eine gelingende Operationalisierung gewährleisten zu können, muss das dynamische Assessment im Laufe der Testplanung nun noch weiter eingegrenzt und präzisiert werden. Für eine solche Eingrenzung bietet es sich in einem ersten Schritt an, die unterschiedlichen Arten des dynamischen Assessments genauer zu betrachten und voneinander abzugrenzen.

2.4 Systematiken des dynamischen Assessments

Da das dynamische Assessment als Sammelbegriff zu verstehen ist (Häcker, 2013), gibt es trotz ähnlicher historischer Wurzeln zwischen den einzelnen Arten des dynamischen Assessments teilweise große Unterschiede. Diese sollen nachfolgend für zwei Klassifikationssysteme dargelegt werden. Zunächst soll das interventionistische Testformat dem interaktionistischen Ansatz gegenübergestellt werden. Anschließend werden *test-train-tests* und *train-within-tests* definiert und voneinander abgegrenzt.

2.4.1 Interaktionistischer versus interventionistischer Ansatz

Als Hauptklassifikationskriterium des dynamischen Assessments gilt die Unterscheidung zwischen interaktionistischen (*interactionist*) und interventionistischen (*interventionist*) Ansätzen (z. B. Hassaskhah & Haghparast, 2012).

Interventionistische Ansätze legen den Fokus auf Standardisierung, insbesondere auf standardisierte Hilfestellung und Rückmeldung (Poehner, 2008, S. 44), während interaktionistische Herangehensweise einen interaktiven, qualitativen Ansatz verfolgen (Poehner, 2008, S. 45).

Der interaktionistische Ansatz geht auf die Arbeit von Feuerstein zurück und steht stärker als interventionistische Ansätze in einem Spannungsverhältnis zu heutigen psychometrischen Standards (Poehner, 2008, S. 45). Für die Zielsetzung der hier vorliegenden Arbeit ist das interaktionistische Assessment daher zu vernachlässigen, da der zu konstruierende Lesekompetenztest nach Möglichkeit den aktuellen psychometrischen Standards genügen soll. Es wird aus diesem Grund im Folgenden nur auf das projektrelevante interventionistische Assessment eingegangen werden. Um diese Art des dynamischen Assessments noch genauer zu präzisieren, soll nachfolgend eine weitere relevante Systematisierung beschrieben werden.

2.4.2 *Test-train-tests* versus *train-within-tests*

Dynamisches Assessment lässt sich in zwei für dieses Projekt relevante Subgruppen untergliedern: *Test-train-tests* und *train-within-tests* (vgl. Dörfler, Golke & Artelt, 2009). Der Fokus des *test-train-test*-Formats liegt auf dem Gedanken der Förderung. Abbildung 1 zeigt den Ablauf einer solchen Testung. Zunächst wird ein Prätest durchgeführt, an den sich eine oder mehrere Trainingssitzungen anschließen. Nach Beendigung des Trainings findet ein Posttest statt. Das Ausmaß, in dem ein Proband von dem Training profitiert hat, sollte sich im Leistungsunterschied zwischen Prätest und Posttest zeigen. Nach Dörfler et al. (2009) werden *test-train-test*-Designs häufiger in der psychologischen Praxis angewandt, sind jedoch als weniger ökonomisch anzusehen.

Abbildung 1: Dynamisches Assessment: Schematischer Ablauf einer dynamischen Testung im test-train-test-Format

Der Fokus des *train-within-test*-Formats liegt auf dem Gedanken der validen Diagnostik. Daher ist die Testung anders organisiert. Es gibt nur eine Testsitzung, die Intervention ist in die Testung integriert. Beckmann und Guthke (1999) sprechen in diesem Zusammenhang auch von sogenannten Kurzzeit-Lerntests in Abgrenzung zu den Langzeit-Lerntests, Sternberg und Grigorenko (2002) vom *cake format* in Abgrenzung zum *sandwich format*. Mischformen zwischen Kurz- und Langzeit-Lerntests sind möglich (Guthke & Wiedl, 1996). Abbildung 2 soll den Ablauf eines *train-within-test*-Formats schematisch darstellen. Der Testand versucht, eine Aufgabe x zu lösen. Gelingt ihm das, so wird ihm die nächste Aufgabe dargeboten. Kann er Aufgabe x nicht lösen, so erhält er eine Hilfestellung und hat einen weiteren Versuch. Die Hilfestellung kann aus Lern- und Denkhilfen oder auch aus fehlerbezogenen Feedbacks bestehen (vgl. Guthke & Wiedl, 1996, S. 77), sie wird standardisiert

gegeben. Es ist prinzipiell möglich, dass ein Testand mehrere Versuche hat, eine Aufgabe richtig zu lösen und bei mehreren falschen Antworten daher mehr als einmal eine Hilfestellung erhält (vgl. Kapitel 2.5). Die Stärke der Responsivität des Testanden auf die Hilfestellungen gilt als Indikator für seine Lernfähigkeit.

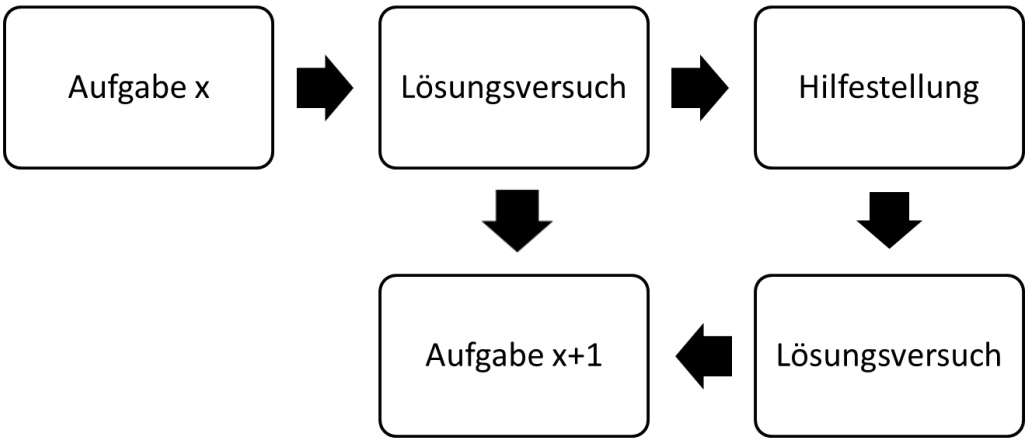

Abbildung 2: Dynamisches Assessment: Schematischer Ablauf einer dynamischen Testung im train-within-test-Format

Diese Unterscheidung hat Implikationen für die durch die dynamische Komponente abzubildende Zone der nächsten Entwicklung. Beckmann und Guthke (1999) halten eine Aufteilung der ZPD in zwei Stufen für sinnvoll. Einen Überblick über die Gestalt der Stufen findet sich beispielsweise bei Baltes und Baltes (1990). Neben der *baseline perfomance*, die im statischen Test erhoben wird, gibt es unter anderem auch die *baseline reserve capacity* und die *developmental reserve capacity*. Mit der *baseline reserve capacity* in einem dynamischen Assessment der Intelligenz ist die zu einem bestimmten Zeitpunkt noch entdeckbare Intelligenzreserve in einer leistungsoptimierenden Testdarbietung gemeint. *Developmental reserve capacity* zielt hingegen auf neue, zusätzliche erzeugbare Kapazitäten ab (vgl. Beckmann & Guthke, 1999). Beckmann und Guthke (1999) gehen davon aus, dass in der Intelligenzdiagnostik mit dem *train-within-test*-Ansatz die Messung der *baseline reserve capacity* und mit dem *train-test-train*-Ansatz die Messung der *developmental reserve capacity* gelingt. Damit sind beide Ansätze als sich

gegenseitig ergänzend anzusehen. Für eine genauere Konstrukteingrenzung im Rahmen der Testplanung müssen diese Ansätze noch weiter spezifiziert werden. Dies soll im Zusammenhang mit der Beschreibung einiger bedeutender interventionistischer Schulen des dynamischen Assessment geschehen, welche als in der aktuellen Forschung relevant angesehen werden können.

2.5 Ausgewählte Ansätze des dynamischen Assessments

Nachfolgend sollen unter besonderer Berücksichtigung deutschsprachiger Ansätze einige besonders verbreitete dynamische Verfahren skizziert werden, die sich als interventionistisch kategorisieren lassen: *Learning potential testing* von Budoff, *graduated prompts* von Campione und Brown, Lerntests von Guthke und *testing-the-limits procedures* von Carlson und Wiedl. Die Darstellungen sollen Basis für eine Identifikation jener Aspekte sein, die für den zu konstruierenden Test Relevanz haben.

Bei den Ansätzen von Guthke sowie von Campione und Brown wird neben den kognitiven Fähigkeiten explizit ein Fokus auf die akademischen Fähigkeiten gelegt (Grigorenko, 2009). Obgleich bedeutsam, wird in der nachfolgenden Aufführung das *Information-Processing Framework* von Swanson nicht berücksichtigt. Bei diesem steht neben den kognitiven Fähigkeiten primär das Gedächtnis im Fokus, was der Zielsetzung der vorliegenden Arbeit nicht dienlich ist. So ist die Rolle des Gedächtnisses für die Leistung in einem Lesekompetenztest gegenüber der Rolle allgemeiner kognitiver Fähigkeiten von untergeordneter Bedeutung (vgl. Kapitel 3.1.3) Ein kurzer Abriss von Swansons *Information-Processing Framework* findet sich beispielsweise bei Caffrey (2006).

***Learning potential testing* von Budoff**

Budoffs *learning potential testing* ist ein interventionistischer Ansatz mit *test-train-test*-Design (Grigorenko & Sternberg, 1998). Die auf diesem Ansatz beruhenden Tests wurden insbesondere für Kinder mit spezifischem Förderbedarf entwickelt, den Lernern werden standardisierte Hilfestellungen gegeben, die beispielsweise auf essentielle Problemattribute hinweisen (Caffrey, 2006). Damit werden bei der Durchführung mehrere Messzeitpunkte realisiert: ein der Intervention vorgeschalteter Prätest, das Training und der nach der Intervention angesetzte Posttest.

Durch einen Vergleich der Leistung im Prätest mit der Leistung im Posttest kann ermittelt werden, in welchem Ausmaß der Lerner von dem Training

profitiert hat (*learning potential*). So können verschiedene Gruppen von Lernern ausgemacht werden: *high scorers*, die gute Leistungen im Prätest und im Posttest zeigen, *gainer*, die sich vom Prätest zum Posttest verbessern und *nongainer*, die sowohl im Prätest als auch im Posttest schlechte Leistung erbringen (Poehner, 2008).

Budoffs Ansatz ist durch eine hohe Standardisierung gekennzeichnet und lässt sich auch an sehr großen Gruppen von Lernern umsetzen (Poehner, 2008). Er kann trotz fehlendem theoretischen Fundament als empirisch abgesichert gelten (Grigorenko & Sternberg, 1998). Testverfahren, die auf Budoffs Ansatz beruhen, sind beispielsweise *Raven's Learning Potential Test* (RLPT), *the Picture Word Game* (PWG), *the Series Learning Potential Test* (SLPT) und *Kohs Learning Potential Test* (KLPT) (Caledo & Márquez, 1998).

***Graduated Prompts* von Campione und Brown**

Bei diesem auch als *testing through learning and transfer* (Caffrey, 2006) bekannten Ansatz wird dem Lerner mehrfach Rückmeldung gegeben. Diese ist standardisiert und in ihrer Reihenfolge determiniert: Zunächst werden sehr unbestimmte Hilfestellungen gegeben. Wenn diese nicht zum Erfolg führen, so werden die Hilfestellungen immer expliziter. Die Anzahl an benötigten Hilfestellungen ist als eine Art Lerngeschwindigkeit zu verstehen und soll dabei jene Dimension erfassen, die der Lernfähigkeit im Sinne Wygotskis (vgl. Kapitel 2.1 und Kapitel 2.2) entspricht. Können die Aufgaben schließlich selbstständig gelöst werden, so wird kein Posttest im Sinne einer einfachen Testwiederholung umgesetzt, sondern Transferaufgaben dargeboten. Diese orientieren sich zunächst inhaltlich nah an den Aufgaben, bei denen der Lerner Hilfestellungen bekommen hat (*near transfer*). Die anschließenden *far transfer*-Aufgaben sind inhaltlich schon weiter von den gelernten Regeln und Prinzipien entfernt und erfordern ein höheres Maß an Transferleistung. Damit wird neben der Schnelligkeit, mit der gelernt wird auch erfasst, wie weit das Gelernte auf neuartige Probleme transferiert werden kann (Poehner, 2008).

Lerntests von Guthke

Die an der Universität Leipzig entwickelten Lerntests (Leipziger Lerntest) sind *within-test*-Verfahren. Ausgehend vom Ansatz der Zone der nächsten Entwicklung wird angenommen, dass es für verschiedene Kompetenzen verschiedene Zonen der nächsten Entwicklung gibt, die mit verschiedenen Tests untersucht werden (Poehner, 2008).

Die bei falschen Antworten gegebene Hilfestellung ist – ähnlich wie beim *testing through learning and transfer* zunächst ein sehr vage und wird immer expliziter. Es werden allerdings höchstens fünf Versuche pro Aufgabe zugelassen. Ist der letzte Versuch falsch, so wird dem Lerner die richtige Lösung mit einer Begründung genannt. Die Anzahl an Versuchen und die für den Test benötigte Zeit werden ebenso erhoben wie die Art der Fehler, die der Lerner machte und die Intervention auf die der Lerner besonders gut ansprach. Auf Basis dieser Daten können für den jeweiligen Lerner Profile erstellt werden, die als Ausgangspunkt für weitere Fördermaßnahmen dienen. Dies erweitert den Ansatz von Budoff, bei dem der Lerner nur klassifiziert wird (Poehner, 2008).

***Testing-the-limits procedures* von Carlson und Wiedl**

Der Ansatz von Carlson und Wiedl bedient sich der *information-processing theory*. Die Performanz in einem Test wird hierbei als ein Zusammenspiel aus dem einzelnen Testanden, dem Testmaterial und der Testsituation gesehen, wobei der Fokus auf der Testsituation liegt, die vom Testleiter so verändert werden kann, dass sie bei Testanden mit Lernproblemen leistungsfördernd wirkt (Caffrey, 2006). Im Gegensatz zum Ansatz von Guthke werden nach richtigen und falschen Antworten Nachfragen gestellt. Der Testand verbalisiert somit seine Gedankengänge. Diese zusätzliche Information trägt dazu bei, die Lernfähigkeit des Testanden besser erfassen zu können, als die Herangehensweisen von Budoff oder Guthke es vermögen (Poehner, 2008).

Empirische Befunde

In empirischen Studien konnten sich - insbesondere auch im deutschsprachigen Raum - diese Ansätze und die darauf basierenden Testverfahren bewähren (vgl.

z. B. Hippmann, 2008; Beckmann, 2004; Guthke, Klauer & Vahle, 2002; Wolschke, Wilmes, Huber & Guthke, 1995; Stein, 1993; Carlson & Wiedl, 1992; Guthke, 1992; Guthke, Wolschke, Willmes & Huber, 1992; Ferrara, Brown & Campione, 1986; Campione & Brown, 1985; Carlson & Wiedl, 1979; Budoff & Corman, 1976). Kein Verfahren kann dabei generell als den anderen Verfahren überlegen angesehen werden. Eine vergleichende Bewertung ist daher nur in Abhängigkeit von spezifischen Fragestellungen und daraus abgeleiteten spezifischen Bewertungskategorien sinnvoll. Inwieweit welche Verfahren für den zu konstruierenden dynamischen Lesekompetenztest als besonders zielführend zu bewerten sind, soll nachfolgend eruiert werden.

Vergleichende Würdigung in Hinblick auf die angestrebte Test-konstruktion

Anhand der vorgestellten theoretischen Ansätze wird deutlich, dass die dynamische Komponente des dynamischen Assessments unterschiedlich operationalisiert wird. Welche Art der Erfassung der dynamischen Komponente besonders exakte und valide Ergebnisse liefert, ist noch nicht abschließend geklärt (Hurley & Murphy, 2015). Sicher ist dies auch dem Umstand geschuldet, dass die Auswirkungen einer Intervention auf den Lerner komplex und von vielen Faktoren abhängig sind (Compton, 2006).

Für die angestrebte Testkonstruktion ist somit eine vergleichende Bewertung sinnvoll, um die Aspekte der einzelnen Ansätze identifizieren zu können, die nach Möglichkeit Eingang in den zu entwickelnden dynamischen Lesekompetenztest finden sollen. Von besonderer Bedeutung für diese Bewertung sind neben Überlegungen zur Objektivität des Tests insbesondere auch seine Ökonomie und Zumutbarkeit. Sie sind valide Kriterien, nach denen die einzelnen Ansätze an Hand der Beschreibungen miteinander verglichen werden können. Daneben kann der Grad ihrer Umsetzung als Marker der Qualität eines Tests angesehen werden, da es sich um Testgütekriterien im Sinne der psychologischen Diagnostik handelt (Pospeschill, 2010).

Ein Testverfahren ist dann als ökonomisch anzusehen, wenn es „gemessen am diagnostischen Erkenntnisgewinn, relativ wenig finanzielle und zeitliche

Ressourcen beansprucht“ (Moosbrugger & Kelava, 2012, S. 21). Zumutbarkeit ist gegeben, wenn der Test „absolut und relativ zu dem Nutzen die zu testende Person in zeitlicher, psychischer sowie körperlicher Hinsicht nicht über Gebühr belastet“ (Moosbrugger & Kelava, 2012, S. 22). Beide Forderungen sind in diesem Zusammenhang zueinander zielkongruent, bestimmte Maßnahmen zur Förderung der Zumutbarkeit wirken sich positiv auf gewisse Ökonomieaspekte aus. Dem Aspekt der Zumutbarkeit kommt insbesondere im Hinblick auf die angestrebte Zielpopulation der Kinder mit spezifischem Förderbedarf eine wichtige Bedeutung zu. Es kann davon ausgegangen werden, dass Probanden mit spezifischem Förderbedarf im Bereich des Lesens in einem Lesekompetenztest besonders stark von den Anforderungen des Tests beansprucht werden und damit schneller ihre Belastungsgrenze erreichen. Um die Belastung nicht unverhältnismäßig stark werden zu lassen, muss neben inhaltlichen Aspekten wie dem Schwierigkeitsniveau der einzelnen Testaufgaben auch auf eine angemessene Testdauer zu achten sein. Sie soll möglichst kurz sein, was sich auch positiv auf die Motivation der Kinder auswirken soll.

Der Ansatz von Budoff ist durch seine empirische Absicherung und seine hohe Standardisierung im Grunde genommen gut für die Anwendung bei der Konstruktion des dynamischen Lesekompetenztests geeignet. Lediglich das aufwändige *test-train-test*-Design müsste modifiziert werden. Unter Berücksichtigung der Gesichtspunkte der Zumutbarkeit und Ökonomie sind *train-within-test*-Designs den *test-train-test*-Designs vorzuziehen, da sie innerhalb einer einzigen Sitzung durchzuführen sind. Damit ist neben einer guten Durchführungsökonomie auch eine bessere Konstruktions- und Auswertungsökonomie gegeben, da nur für eine Sitzung Materialien erstellt und erhobene Testwerte ausgewertet werden müssen. Dies ist in diesem Maße beim klassischem Ansatz nach Budoff weniger gegeben, bei dem zwei Testungen durchgeführt werden, denen eine Intervention zwischengeschaltet ist. Darüber hinaus kann der auf der besonderen Validität der Diagnostik liegende Fokus des *train-within-test*-Formats als positiv gewertet werden.

Die Umsetzung des *train-within-test*-Formats findet sich beispielsweise bei den Ansätzen von Guthke sowie von Campione und Brown wieder. Beide verwenden innerhalb einer Testung Rückmeldungen, nach einer falschen Antwort kann der Lerner die Aufgabe erneut bearbeiten. Die Anzahl an Feedbackschleifen und damit auch die Gesamtzahl der möglichen Versuche pro Aufgabe ist beim Leipziger Lerntest begrenzt, Campione und Brown setzen dagegen zunächst einmal keine solche Begrenzung. Stattdessen wird mit den Transferaufgaben eine weitere Dimension in die Testung eingeführt. Deren Sinnhaftigkeit wird nicht in Frage gestellt, jedoch stehen solch zusätzliche Transferaufgaben im Anschluss an die dynamische Testung in einem Spannungsverhältnis zu den angestrebten Zielen der Zumutbarkeit und der Ökonomie. So stellen die Transferaufgaben nach der Testung einen zeitlichen und kognitiven Mehraufwand und damit eine zusätzliche Belastung für die Probanden dar und erhöhen die für die Konzeption, Durchführung und Auswertung der Testung benötigten Ressourcen. Hier entspricht der Ansatz von Guthke diesen Zielen besser.

Weniger geeignet für eine Umsetzung im zu konstruierenden Lesekompetenztest ist der Ansatz von Carlson und Wiedl. Sicher bildet ihr Ansatz, bei dem ein Zusammenspiel aus dem einzelnen Testanden, dem Testmaterial und der Testsituation, auf der das besondere Augenmerk liegt, die der Testleistung zu Grunde liegende Komplexität gut ab. Gleichzeitig erschwert diese Komplexität die Spezifizierung und Objektivierung der zu erfassenden Konstrukte. Auch sind die spezifischen Nachfragen und die Verbalisierung der Gedankengänge des Lerners nicht problemlos quantifizierbar. Daher sind an dieser Stelle andere Ansätze der dynamischen Testung zu präferieren.

Empirische Befunde legen nahe, dass auch ein Design eine valide dynamische Testung umsetzen kann, welches noch ökonomischer als der Leipziger Lerntest ist. Hierbei wird nur eine Hilfestellung nach einer falschen Antwort und damit nur zwei Versuche pro Aufgabe zugelassen (vgl. Golke et al., 2015). Inhaltlich ist diese Herangehensweise in gewisser Nähe zu Budoff anzusiedeln, bei dem die Performanz vor und nach der Intervention miteinander verglichen wird. Es

umgeht die mangelnde Ökonomie des *test-train-test*-Designs, indem es dieses Design in die Testung selbst verlagert und zwar auf die Ebene des Items. Jedes Item stellt damit für sich genommen eine Testung im *test-train-test*-Design da. Gleichzeitig werden alle Items innerhalb einer einzigen Testsitzung administriert, was die Belastung für die Probanden reduziert und die Konstruktions-, die Durchführungs- und die Auswertungsökonomie verbessert. Dieses Vorgehen soll daher im zu konstruierenden dynamischen Lesekompetenztest umgesetzt werden.

Eine weitere Verbesserung der Ökonomie kann durch eine Maßnahme erzielt werden, die insbesondere auch der Objektivität des Tests zu Gute kommt: die computeradministrierte Umsetzung des dynamischen Lesekompetenztests.

Durch eine Testung am Computer kann die Testdurchführung und -auswertung als standardisiert und besonders unabhängig von der Person des Testleiters angesehen werden, was einem Merkmal des Gütekriteriums der Objektivität entspricht (Pospeschill & Spinath, 2009, S. 57). Während eine computeradministrierte Testung sich beispielsweise beim Ansatz von Carlson und Wiedl nicht problemlos implementieren lässt, so sollte sie im hier angestrebten Design gut umsetzbar sein. Innerhalb einer PC-Testung lassen sich Aufgaben, Antworten der Probanden und Rückmeldungen mit Testaufgaben beispielsweise im Multiple-Choice-Format besonders leicht standardisiert erheben. In empirischen Studien hat sich gezeigt, dass das dynamische Assessment computeradministriert durchgeführt werden kann (vgl. Kapitel 2.3).

Neben diesen technischen Überlegungen ist es für den zu konstruierenden Test jedoch auch noch nötig, spezifische inhaltliche Problembereiche zu berücksichtigen, die bei der Operationalisierung der Lernfähigkeit auftreten können. Diese sollen nachfolgend skizziert werden.

2.6 Psychometrische Kritikpunkte

In Bezug auf die Messung der Lernfähigkeit sind insbesondere drei psychometrische Problembereiche nach Sternberg und Grigorenko (2002) relevant:

1. *Gain scores* sind nicht als reliabel anzusehen, denn die Differenzen zwischen Prä- und Posttest werden reliabler, wenn die Korrelation zwischen Prä- und Posttest abnimmt, d. h. wenn mindestens eine dieser beiden Messungen weniger reliabel wird.

2. Es können Deckeneffekte auftreten. Testanden, die bereits im Prätest hinreichend gut sind, werden sich durch die Interventionen nicht mehr stark verbessern können und müssen bei Analysen gesondert betrachtet werden.

3. Die Skala […] ist nicht hinreichend gut verstanden. Klar ist jedoch, dass sie keine Äquidistanz aufweist.

Darüber hinaus sind dynamische Verfahren besonders zeit- und kostenintensiv und weniger stark als konventionelles Assessment durch theoretische Modelle und Verbindungen zu anderen psychologischen Disziplinen legitimiert (Murphy, 2011).

Während diese Kritikpunkte nicht generell das Konzept des dynamischen Assessments ad absurdum führen, so sind sie doch bei der Entwicklung, Anwendung und Evaluation dynamischer Tests angemessen zu berücksichtigen und sollen auch in der vorliegenden Arbeit in den folgenden Kapiteln, in denen sie von Relevanz sind, angesprochen werden.

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Spezifisch folgt aus den psychometrischen Kritikpunkten für das zu konstruierende Testverfahren, dass Differenzen zwischen Prä- und Posttest als problematisches Maß angesehen werden können. Ob sich Differenzen als

spezifisches Maß vermeiden lassen, ist vom konkreten Einzelfall abhängig. Jedoch sollten unabhängig vom konkret verwendeten Maß a priori möglichst wenige Vorannahmen über die Skala der Lernfähigkeit und ihre spezifischen Eigenschaften getroffen werden. Dagegen kann das Problem der Deckeneffekte nicht vom Kontext der Stichprobenbeschaffenheit isoliert betrachtet werden. Es wird erwartet, dass Deckeneffekte insbesondere in der Population der Kinder mit spezifischem Förderbedarf keine herausragende Bedeutung haben werden. Außerdem kann durch die Entwicklung eines möglichst ökonomischen Testverfahrens dem Problem der Zeit- und Kostenintensivität begegnet werden (vgl. Kapitel 2.6) und die Verbindungen zu anderen Disziplinen gestärkt werden.

Da die Problemfelder der dynamischen Testung meist im Vergleich mit der statischen Testung deutlich werden, liegt der Gedanke nahe, statische und dynamische Testungen als „Rivalen“ wahrzunehmen. Sicher ist es jedoch zu kurz gegriffen, dynamisches Assessment lediglich als „Ersatz“ für statische Verfahren zu verstehen und diese beiden Ansätze direkt in Konkurrenz zueinander zu setzen. Dies soll nachfolgend am Beispiel der dynamischen Testung der Intelligenz noch einmal genauer dargestellt werden.

2.7 Dynamisches Assessment im Kontext der kanonischen Intelligenzforschung

Während Lernpotential heute auch als Teil der allgemeinen Intelligenz aufgefasst werden kann (Sarges, 2013) ist Lernpotential im Sinne von Wygotski von diesem Intelligenzbegriff abzugrenzen. In der sowjetischen Psychologie war dieser Intelligenzbegriff und die darauf aufbauende Intelligenzdiagnostik politisch nicht erwünscht (Grigorenko, 2004, S. 178). Sowjetischen Psychologen verwendeten dieses westliche Konzept nicht (Grigorenko, 2004, S. 204), ihre Testungen der intellektuellen Entwicklung unterschieden sich von der westlicher Länder in ihrer Konzeptualisierung, in ihrer Verbreitung und in der Akzeptanz, die dem theoretischen Begriff der „Intelligenz“ entgegengebracht wurde (Grigorenko, 2004, S. 179). Daher kann Lernpotential als ein von der allgemeinen Intelligenz nach unserem kulturellen Begriff distinktes Konzept angesehen werden (Grigorenko, 2004, S. 177). Die Fähigkeit zu Lernen steht jedoch in einem Zusammenhang mit Intelligenz (Amelang, Bartussek, Stemmler & Hagemann, 2006, S. 205) und kognitiven Prozessen (z. B. Kovalčíková, 2015) und es besteht Grund zu der Annahme, dass die der Intelligenz und dem Lernpotential zu Grunde liegenden kognitiven Prozesse dieselben sind (Jensen, 1989).

Diese Abgrenzung und Unterscheidung vom Konzept der Intelligenz wird auch nicht durch die Ansicht in Frage gestellt, dass Intelligenz als *ability to learn* charakterisiert werden kann (z. B. Sternberg, 2000; Brown, Campione, Webber & McGilly, 1992, S. 130). Obgleich das Konzept der Lernfähigkeit damit zwar auf den ersten Blick eine starke inhaltliche Nähe zur Intelligenz aufzuweisen scheint, so ist sie doch unter anderem durch ihren theoretischen Hintergrund (Kapitel 2.2) vom Konzept der Intelligenz hinreichend distinkt.

Dies impliziert, dass Maße, die die Lernfähigkeit operationalisieren sollen, demnach keine sehr hohen sondern eher moderate Korrelationen mit Intelligenzmaßen aufweisen sollen. Die exakten Zusammenhänge sind unter anderem von der Art und Ausrichtung des dynamischen Assessments abhängig.

Die Frage, ob Lerntests eine „bessere Version“ des Intelligenztests darstellen, wie sie Guthke und Stein (1996) aufwarfen, ist nach diesem Verständnis so nicht zu beantworten, da es sich bei Lerntests nicht um eine Version des Intelligenztests handelt, sondern um Testverfahren, die auf ein anderes Konstrukt als die Intelligenz abzielen. Empirisch lässt sich dies durch Befunde untermauern, die aufzeigen, dass niedrige Intelligenz durchaus mit hoher Lernfähigkeit einhergehen kann (Brown & Ferrara, 1999) und der im hohen Maße intelligenzabhängige schulische Erfolg (Amelang et al., 2006, S. 205) kaum im Zusammenhang mit der Performanz in einem dynamischen Test steht (Guthke, 1992).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

In diesem Sinne ist auch in der vorliegenden Arbeit die Lernfähigkeit als ein Konstrukt zu verstehen, welches nicht mit Intelligenz identisch ist, aber positiv mit kognitiven Maßen korrelieren sollte. Es zielt auf Veränderungen ab, die durch Interventionen im Rahmen der Testung induziert werden können. Die Zusammenhänge mit intelligenznahen Konstrukten sollten eine moderate Höhe aufweisen. Zusammenhänge mit Maßen des schulischen Erfolgs sind dagegen kaum gegeben.

2.8 Zusammenfassung und Implikationen für die Testkonstruktion

Bei dynamischen Verfahren kann von einer Dichotomie der zu erfassenden Konstrukte gesprochen werden, neben dem interessierenden Merkmal wird auch seine gezielt evozierte intraindividuelle Variabilität (Lernfähigkeitskomponente) erhoben. Dies setzt voraus, dass das interessierende Merkmal gezielt im positiven Sinne verändert werden kann. Ein solches Verfahren kann beispielsweise konstruiert werden, indem eine zunächst statische Testversion um eine dynamische Komponente erweitert wird. (Kapitel 2.1).

Empirische Befunde legen nahe, dass die computeradministrierte Umsetzung eines dynamischen Tests bei Schülern der Jahrgangsstufen 3 bis 4 unter besonderer Berücksichtigung der Kinder mit spezifischem Förderbedarf möglich ist. Geschlechtsspezifische Auswirkungen der Interventionen im Rahmen dynamischer Testverfahren sind dabei in Anlehnung an Ramazanpour et al. (2016) nicht zu erwarten. Der Gedanke der Computertestung trägt daneben auch dem Streben nach einer möglichst objektiven und ökonomischen Testung Rechnung, dem Kritikpunkt der im Vergleich zu statischen Verfahren weniger gegebenen Testökonomie (Kapitel 2.6) kann damit teilweise entgegengewirkt werden. Neben den empirischen Befunden (Kapitel 2.3) spricht auch das theoretische Fundament des dynamischen Assessments (Kapitel 2.2) für eine hohe Wahrscheinlichkeit, einen validen dynamischen Lesekompetenztest konstruieren und validieren zu können.

Da das Konzept der Lernfähigkeit per se noch nicht vollständig geklärt ist, kann das Ausmaß, in welchem die Schüler auf die ihnen gegebenen Rückmeldungen ansprechen, in dieser Arbeit auf eine neue Art und Weise operationalisiert werden, sodass den Spezifika des Projekts im Besonderem Rechnung getragen wird. Ihre exakte Herleitung unter Berücksichtigung der in diesem Kapitel aufgeführten Aspekte findet sich in Kapitel 8.3.2.3. Aus diesem Grund können a priori wenige Vorannahmen über ihre Korrelationen mit anderen Variablen getroffen werden. Es ist jedoch von einem positiven Zusammenhang mit allgemeinen kognitiven Fähigkeiten und intelligenznahen

Maßen auszugehen, wobei die Korrelationen in ihrer Stärke moderat ausfallen sollten. Dieser ist in Teilen der Tatsache zuzuschreiben, dass die Intelligenz und die Lernfähigkeit auf unterschiedlichen theoretischen Fundamenten begründet sind. Empirische Befunde zum mangelnden Zusammenhang zwischen Lernfähigkeit und Schulerfolgsindikatoren wie dem Lehrerurteil oder den Schulnoten (Guthke, 1992) stützen diese Annahme (Kapitel 2.7).

Bei der Operationalisierung der dynamischen Komponente sollten aus psychometrischen Gesichtspunkten a priori generell möglichst wenig Vorannahmen über die Skala und ihre spezifischen Eigenschaften getroffen werden. Darüber hinaus sind Differenzen zwischen Prä- und Posttestung als Maß der Lernfähigkeit problematisch. Das Problem der Deckeneffekte kann dagegen insbesondere in der Population der Kinder mit spezifischem Förderbedarf als weniger bedeutsam angesehen werden.

Psychometrische Gründe sprechen auch für den interventionistischen Testansatz (Kapitel 2.4.1), der mit Blick auf die Ökonomie und die Zumutbarkeit des Testverfahrens im *train-within-test*-Design umgesetzt werden soll (vgl. Kapitel 2.4.2). Dabei soll jeweils ein standardisiertes Feedback auf eine falsche Antwort gegeben werden, bei jeder Testaufgabe hat der Proband nach einer falschen Antwort einen zweiten Versuch (vgl. Kapitel 2.5).

Der Inhalt der standardisierten Rückmeldung muss auf den Kompetenzbereich Lesen abzielen, der auch der Gegenstand der eigentlichen Testaufgaben ist. Damit wird ein zunächst statischer Lesekompetenztest entwickelt und um eine dynamische Variante ergänzt. Die konkrete inhaltliche Ausrichtung dieses statischen Lesekompetenztests soll nachfolgend hergeleitet werden.

3 Lesekompetenz

3.1 Definition, Komponenten und Einflussfaktoren

3.1.1 Begriffliche Eingrenzung

Ein wesentlicher Gegenstand dieser Arbeit ist die Lesekompetenz. Lesekompetenz ist als komplex anzusehen, mehrere Faktoren greifen beim Lesen ineinander (z. B. Groeben & Christmann, 2013; Rosebrock & Nix, 2013; Klicpera & Gasteiger-Klicpera, 1998). Der Begriff der Lesekompetenz und seine Komponenten sind nicht eindeutig definiert, sollen aber nachfolgend präzisiert und dabei die Komponenten kurz umrissen werden. Dabei ist das konzeptuelle Verständnis der Lesekompetenz aus den Konzepten seiner Teilaspekte Kompetenz und Lesen nachvollziehbar.

Kompetenz kann verstanden werden als die „bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“ (Weinert, 2001, S. 27f.). Der Kern des Konstrukts der Kompetenz ist nach dieser Definition kognitiv und Kompetenz kann „als Befähigung zur Bewältigung unterschiedlicher Situationen“ (Klime, 2004) aufgefasst werden.

Lesen ist als „ein eigenaktiver Prozess der Sinnkonstruktion“ (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2005) zu verstehen, bei der die Leser neben einer visuell orientierten Verarbeitungsweise zur Dekodierung graphischer Zeichen auch sprachlich formale Kompetenzen und bereichsspezifisches Vorwissen benötigen (Nold & Willenberg, 2007).

Die Verbindung des Begriffs der Kompetenz mit dem des Lesens findet sich beispielsweise in der Definition der Lesekompetenz der Internationalen Grundschul-Lese-Untersuchung (IGLU), die unter Lesekompetenz die

Fähigkeit versteht, „Lesen in unterschiedlichen, für die Lebensbewältigung praktisch bedeutsamen Verwendungssituationen einsetzen zu können“ (Bos, Lankes, Prenzel et al., 2003, S. 73). Die für die Lesekompetenzerhebung im Rahmen der an Schülern der Sekundarstufe I durchgeführten PISA-Studien gültige Definition lautet:

> Unter Lesekompetenz versteht PISA die Fähigkeit, geschriebene Texte unterschiedlicher Art in ihren Aussagen, ihren Absichten und ihrer formalen Struktur zu verstehen und in einen größeren Zusammenhang einordnen zu können, sowie in der Lage zu sein, Texte für verschiedene Zwecke sachgerecht zu nutzen. Nach diesem Verständnis ist Lesekompetenz nicht nur ein wichtiges Hilfsmittel für das Erreichen persönlicher Ziele, sondern eine Bedingung für die Weiterentwicklung des eigenen Wissens und der eigenen Fähigkeiten – also jeder Art selbstständigen Lernens – und eine Voraussetzung für die Teilnahme am gesellschaftlichen Leben. (Artelt, Baumert et al., 2001, S. 11)

Insgesamt ist damit in beiden Untersuchungen der Fokus der Lesekompetenz auf die Alltagstauglichkeit und die Relevanz des Lesens für das tägliche Leben des Lesers gelegt. Der in der PISA-Studie etwas anders gesetzte Schwerpunkt bei der Definition der Lesekompetenz ist unter anderem auf ein anderes Alter der Zielpopulation zurückzuführen, bei denen grundlegende Lesetechniken weniger im Vordergrund stehen als bei der IGLU-Studie (Artelt, Drechsel, Bos & Stubbe, 2008, S. 39).

Lesekompetenz zielt damit insbesondere auf das sinnentnehmende und -verstehende Lesen von Texten ab. Damit dieses funktioniert, müssen beim Lesen mehrere Prozesse gleichzeitig ablaufen und ineinandergreifen.

3.1.2 Prozesse des Lesens

Die beim Lesen ablaufenden Prozesse können auf unterschiedliche Art und Weise beschrieben werden. So lassen sich nach Richter und Christmann (2002) hierarchiehohe und hierarchieniedrige Leseprozesse ausmachen. Hierarchiehöhere Prozesse stellen an den Rezipienten mehr Anforderungen kognitiver Art, während hierarchieniedrigere Prozesse eher automatisch ablaufen. Emotionale und motivationale Gesichtspunkte bleiben bei diesem Ansatz unberücksichtigt (Richter & Christmann, 2002).

Der Prozess des Lesens umfasst verschiedene Verarbeitungsschritte auf Ebene der Wörter, Sätze und Texte (Christmann & Groeben, 1999; Klicpera & Gasteiger-Klicpera, 1998; Hohm, 2005). Daneben wird auch das Erkennen von Buchstaben als eine Komponente des Lesens erachtet (Artelt, Stanat, Schneider & Schiefele, 2001; Lenhard & Artelt, 2009). Sie wird von Christmann und Groeben (1999) zur Wortebene gezählt. Lenhard und Artelt (2009) teilen die beim Lesen ablaufenden Prozesse in die Bereiche Vorläuferfertigkeiten, hierarchieniedrige Prozesse auf Wort- und Satzebene, satzübergreifendes Lesen und Textverständnis ein. Einen groben Überblick über die beim Lesen relevanten Ebenen ist in Tabelle 1 dargestellt.

Tabelle 1: Übersicht über verschiedene Leseprozesse

Spezifische Ebene	**Art der Prozesse**
Vorläuferfertigkeiten	-
Buchstabenebene	Hierarchieniedrig
Wortebene	Hierarchieniedrig
Satzebene	Hierarchieniedrig
Textebene	Hierarchiehoch

Beim Identifizieren von Buchstaben und Wörtern ist von primär visuellen Verarbeitungsvorgängen auszugehen, hohe kognitive Anforderungen an den Leser spielen hier noch eine relativ kleine Rolle. Bei den Prozessen auf Satzebene müssen die identifizierten Wörter unter Berücksichtigung von

Syntax und Semantik derart aufeinander bezogen werden, dass koheränte Sinnstrukturen aufgebaut werden. Mit diesem Schritt steht der Prozess des Verstehens erst am Anfang, die Bedeutungseinheiten auf Satzebene müssen nun in ein sinnvolles, kohärentes Gesamtgefüge auf Textebene gebracht werden. Dafür müssen die im Text gegebenen Informationen um eigene Schlussfolgerungen ergänzt werden (Christmann & Groeben, 1999). Dieser als Inferenzbildung bekannte Vorgang spielt eine wichtige Rolle beim verstehenden Lesen und stellt vergleichsweise hohe kognitive Anforderungen an den Leser. Er wird in Kapitel 3.2 nochmal genauer ausgeführt.

Man geht davon aus, dass die Prozesse auf Wort-, Satz- und Textebene nicht sequentiell ablaufen, sondern sich vielmehr zeitlich überlappen. Gleichzeitig sind sie voneinander abhängig, hierarchiehöhere Prozesse basieren auf den Ergebnissen hierarchieniedrigerer Prozesse (Christmann & Groeben, 1999).

Ein wichtiges Fundament dieser Prozesse bilden die Vorläuferfähigkeiten des Lesens. Als eine für die Entwicklung der Lesekompetenz bedeutende Vorläuferfähigkeit, die sich bereits im Vorschulalter ausbildet, wird insbesondere die phonologische Bewusstheit angesehen (Hatz, 2015; von Goldammer, 2010; Klicpera & Gasteiger-Klicpera, 1998; Elbro, 1996). Sie kann verstanden werden als das „Bewusstsein, dass Wörter mit ihrer jeweiligen Klanggestalt isolierbar sind und nicht mit der Sache, die sie bedeuten, in eins fallen“ (Rosebrock & Nix, 2013).

Die kausale Bedeutung der Vorläuferfähigkeiten für die gelingende Entwicklung der Lesekompetenz wird von metaanalytischen Befunden unterstützt (Pfost, 2015; Melby-Lervåg, Lyster & Hulme, 2012). Hierbei fanden sich beispielsweise bei 21 Studien aus dem deutschsprachigen Raum, die den Zusammenhang der phonologischen Bewusstheit mit dem Schriftspracherwerb eruierten, durchschnittliche Effektstärken von Z_r= 0.318 (r=.308) (Pfost, 2015).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Beim sinnentnehmenden und -verstehenden Lesen von Texten ist es im Sinne der Vielschichtigkeit des Konzepts sinnvoll, nicht nur ein einziges, eng definiertes Konstrukt zu operationalisieren und zu erfassen. Vielmehr sollte versucht werden, verschiedene Facetten der Lesekompetenz zu berücksichtigen. Die Notwendigkeit für die Erfassung eines heterogenen Lesekompetenzkonstrukts wird nochmal besonders deutlich, wenn man sich die Komplexität des zu messenden Konstrukts vor Augen führt. Nicht nur die einzelnen Teilprozesse greifen beim Lesen ineinander, auch spezifische Merkmale des Lesers und des Textes sind in diesem Zusammenhang von Bedeutung. Nachfolgend sollen ausgewählte Leser- und Textmerkmale dargestellt werden, die im Rahmen dieser Arbeit von Relevanz sind.

3.1.3 Ausgewählte Korrelate der Lesekompetenz

3.1.3.1 Spezifische Merkmale des Lesers

Entsprechend der bislang gemachten Aussagen ist Lesekompetenz als ein multidimensionales Konstrukt zu verstehen, bei dem verschiedene Prozesse zu berücksichtigen sind. Darüber hinaus kommen spezifische Eigenschaften des Lesers und Eigenschaften des Textes zu tragen (z. B. Artelt & Dörfler, 2010; Lenhard & Artelt, 2009; Ozuru, Dempsey & McNamara, 2009; Hurrelmann, 2002). Diese kommen nicht isoliert zum Tragen, sondern stehen miteinander in Wechselwirkung (z. B. McNamara, 2001). Begründen lassen sich diese Zusammenhänge in der Definition des Lesens als „aktiver Prozess der Bedeutungskonstruktion […], bei dem die Leser die Textinformation mit ihrem Vor- wie Weltwissen verbinden“ (Christmann & Groeben, 2013, S. 960). Die für diese Arbeit besonders relevanten Merkmale des Lesers sind die allgemeinen kognitiven Fähigkeiten, das Vorwissen und die (Test-)ängstlichkeit des Lesers sowie sein Geschlecht. Demografische Angaben wie die aktuell besuchte Schulart und Klassenstufe eines Kindes lassen ebenfalls Rückschlüsse auf seine spezifische Lesekompetenz zu. Diese Merkmale sollen im Folgenden kurz skizziert werden, bevor die spezifischen Textmerkmale angesprochen werden, die für diese Arbeit von Relevanz sind.

Allgemeine kognitive Fähigkeiten

Es wurde in dieser Arbeit bereits angeschnitten, dass der kognitiven Komponente beim Lesen eine besondere Bedeutung zukommt: Lesen ist ohne kognitive Leistung nicht möglich (Hurrelmann, 2006).

Ein zentraler Begriff stellt hierbei die Fähigkeit zum schlussfolgernden Denken (*reasoning*) da, die charakterisiert werden kann als eine abstrakte und „relativ dekontextualisierte Fähigkeit zum Umgang mit neuen Problemstellungen, für deren Lösung das spezifische Wissen fehlt, also um die Fähigkeit zu Verständnis und Einsicht – zum Erkennen und zur Herstellung von Strukturen, Beziehungen, Sinnzusammenhängen und Bedeutungen“ (Baumert, Brunner, Lüdtke & Trautwein, 2007). Diese Definition bezieht sich damit auf die

„komplexe Fähigkeit zum denkgestützten Lösen von Aufgaben und Problemen in Situationen, die für die Person neu und nicht allein durch Wissensabruf erfolgreich bearbeitbar sind“ (Baumert et al., 2007; Rindermann, 2006) und zielt damit auf die fluide Intelligenz ab (Baumert et al., 2007). Fluide Intelligenz ist kulturunabhängig und umfasst die „Fähigkeit, sich in neuen Situationen orientieren zu können, schlussfolgerndes Denken, Problemlösefähigkeiten sowie die Wahrnehmungsgeschwindigkeit“ (Kessler, 2015, S. 127). Sie wird als Marker des g-Faktors angesehen (Lohaus & Vierhaus, 2015, S. 133) und lässt sich mit figuralen Matrizenaufgaben gut abbilden (Baumert et al., 2007; Preckel, 2002; Hossiep, Turck & Hasella, 2002). Die Verwendung solcher sprachfreier Intelligenztests ist dabei insbesondere bei Kindern mit Defiziten im Lesen und Schreiben sinnvoll, bei denen sprachabhängige Tests zu einer Verzerrung der Testergebnisse führen würden (Lohaus & Vierhaus, 2015, S. 131).

Daneben kann auch das Leseverständnis als „Verarbeitung verbal kodierter Problemstellungen“ (Rost & Schilling, 2006, S. 452; Rost, 1987) angesehen werden. Die Bewertung der Bedeutung der kognitiven Komponente beim Lesen geht teilweise soweit, dass manche Wissenschaftler Lesen im Bereich des (sprachfreien) *reasoning* verorten (Stanovich & Cunningham, 1991). *Reasoning* spielt beim Lesen eine Rolle (z. B. van den Broek, 2010). Auch der g-Faktor lädt auf Leseleistung (Rindermann, 2007). Insgesamt sind die Zusammenhänge der Lesekompetenz mit den allgemeinen kognitiven Fähigkeiten in der einschlägigen Literatur empirisch gut belegt (z. B. Swanson, 2011; Schaffner, 2009; Rost & Sparfeldt, 2007; Harlaar, Hayiou-Thomas & Plomin, 2005; Heinen, 2001; De Jonge & De Jong, 1996).

Die Höhe der Korrelationen zwischen der allgemeinen Intelligenz und den Leseverständnisskalen streuen in der Regel um $r = .60$, teilweise auch darüber (Rost & Sparfeldt, 2007). Der positive Zusammenhang zwischen allgemeiner Intelligenz und Lesen ist überdies zeitlich stabil und bleibt bis ins Erwachsenenalter hinein auffindbar (Johnson, Bouchard, Segal & Samuels, 2005). Daneben ist die Intelligenz von Bedeutung für eine weitere wichtige Einflussvariable der Lesekompetenz, das Vorwissen (Stern, 2001).

Vorwissen

Wie bereits in der Begriffsbestimmung des Lesens in Kapitel 3.1.1 angedeutet, spielt das Vorwissen für das verstehende Lesen eine wichtige Rolle. Es ist gut belegt, dass sich höheres Vorwissen positiv auf die beim Lesen ablaufenden Prozesse auswirkt (z. B. Priebe, Keenan & Miller, 2012; Coiro, 2011; Tarchi, 2010; Ozuru et al., 2009; Taboada, Tonks, Wigfield & Guthrie, 2009; Kendeou & van den Broek, 2007; Artelt, Schiefele & Schneider, 2001; McNamara, 2001). Die Vorkenntnisse des Lesers können „semantischer oder inhaltlicher Natur sein, sie können strukturelle Gegebenheiten eines Textes betreffen; es kann sich aber auch einfach um Alltagswissen handeln, das die Ausbildung von Situationsschemata erleichtert“ (Klicpera, Schabmann & Gasteiger-Klicpera, 2010, S. 75). Insbesondere bei der Rezeption von Sachtexten ist das bereichsspezifische Vorwissen von Bedeutung (Golke, Matthäi & Artelt, 2013; Lenhard, 2013, S. 28).

Testängstlichkeit und allgemeine Ängstlichkeit

Neben diesen Faktoren gibt es auch noch zahlreiche emotionale und motivationale Faktoren, die mit der Leseleistung interagieren, beispielsweise kann Lesekompetenz bzw. die zum Lesen benötigte Zeit von Testängstlichkeit (Calvo & Carreiras, 1993; Gifford & Marston, 1966) oder von allgemein erhöhter Ängstlichkeit (Grills-Taquechel, Fletcher, Vaughn & Stuebing, 2012) beeinflusst sein. Analog dazu kann auch ein genereller Zusammenhang zwischen Leseleistung und spezifischen Persönlichkeitseigenschaften als belegt angenommen werden (Krach, McCreery, Loe & Jones, 2015). Für einen umfassenden Überblick über motivationale Korrelate der Leseleistung sei beispielsweise auf Snow und Verhoeven (2001) verwiesen.

Im Rahmen der vorliegenden Arbeit sind insbesondere die Testangst und die mit ihr oft einhergehende allgemeine Ängstlichkeit (z. B. Moutafi, Furnham & Tsaousis, 2006; Spielberger & Vagg, 1995; Hembree, 1988) von Bedeutung. Testangst kann verstanden werden als Angstgefühl, das im Zusammenhang mit einer Testung auftreten und den wahren Testwert verfälschen kann (*Testangst*, 2013) und kann bereits bei Grundschülern empirisch erfasst und untersucht werden (Harris, 2014; Segool, Carlson, Goforth, von der Embse & Barterian,

2013). Testangst hat negative Auswirkungen auf die akademische Leistung sowie auf die Leistung in Intelligenztests (z. B. Richardson, Abraham & Bond, 2012; Moutafi et al., 2006; Chapell et al., 2005; Cassady & Johnson, 2002) und vermindert spezifisch die Leseleistung (Javanbakht & Hadian, 2014), möglichweise weil sie einhergeht mit spezifischer Angst gegenüber der zu leistenden Aufgabe (Tsai & Li, 2012). So weisen beispielsweise auch Kinder mit *reading disabilities* gegenüber Kindern mit normaler Entwicklung erhöhte Werte in Ängstlichkeit auf (Mammarella et al., 2016) und generell ängstlichere Leser sind weniger effizient beim Lesen (Amelang et al., 2006, S. 373; Calvo & Carreiras, 1993). Dabei kann Ängstlichkeit allgemein als ein „Persönlichkeitskonstrukt, das Unterschiede zwischen Personen hinsichtlich ihrer Wahrscheinlichkeit beschreibt, öfter mit Angst oder mit besonders starken Ängsten zu reagieren“ (Amelang et al., 2006, S. 366) aufgefasst werden. Sie wirkt sich negativ auf akademische Leistung aus (Chamorro-Premuzic & Furnham, 2003). Ihr Zusammenhang mit akademischer Leistung ist laut einer Metaanalyse von Seipp (1991) bei $r=-.163$ und ist damit in ihrer negativen Auswirkung geringer als die in der Metaanalyse gefundene Korrelation der akademischen Leistung mit Testangst, welche bei $r=-.233$ liegt. Damit können allgemeine Ängstlichkeit und spezifische Testangst die Leistung in einem Lesekompetenztest negativ beeinflussen. Dieser Einfluss kann unter anderem auf das Verhalten in der Testsituation zurückgeführt werden: die Aufmerksamkeit wird bei erhöhter Angst von aufgabenrelevanten Stimuli auf Stimuli gelenkt, die mit der Aufgabe nicht in Zusammenhang stehen, wie beispielsweise Reize, die mit der erlebten Bedrohung assoziiert sind. Somit stehen weniger Kapazitäten zur Aufgabenbewältigung zur Verfügung (Eysenck, Derakshan, Santos & Calvo, 2007; Wine, 1971). Hierbei können weitere Faktoren wie emotionale Zustände eine Rolle spielen. Sie können eine Konsequenz des Verhaltens in der Testsituation und der sich daraus ergebenden negativen Leistungsrückmeldung sein und daraufhin in späteren Testsituationen erhöhte Ängstlichkeit mitverursachen (z. B. Steinmayr, Crede, McElvany & Wirthwein, 2015), emotionale und kognitive Faktoren können sich so gegenseitig negativ verstärken.

Die negativen Auswirkungen der Testängstlichkeit sind jedoch auch von der Art der Testung abhängig, so sind sie beispielsweise bei dynamischen Tests weniger stark ausgeprägt als bei konventionellen, statischen Tests (Meijer, 2001; Bethge, Carlson & Wiedl, 1982). Sie können aber auch durch spezifische Maßnahmen innerhalb der Testung reduziert werden, etwa durch einen angstreduzierenden, kindgerechten Umgang der Testleiter mit den Probanden und durch die Zusicherung, dass der Test anonymisiert durchgeführt wird und nicht benotet werden soll, was ihn für die Probanden nicht über die Testsituation hinaus relevant macht.

Geschlecht, Klassenstufe und Schularten

Darüber hinaus sind insbesondere drei demographische Variablen für die kindliche Lesekompetenz von großer Relevanz: das Geschlecht, die Klassenstufe und die besuchte Schulart.

Das Geschlecht gilt als wichtiger Prädiktor der Leseleistung, Mädchen erzielen in hier in empirischen Untersuchungen tendenziell die bessere Ergebnisse (vgl. z. B. Hohn, Schiepe-Tiska, Sälzer & Artelt, 2013; Eckert, 2012; Schabmann, Landerl, Bruneforth & Schmidt, 2012; Bos, Lankes, Schwippert et al., 2003). Dieser Unterschied wird in gängigen psychometrischen Testverfahren in Form von getrennte Normen für Mädchen und Jungen berücksichtigt, so beispielsweise beim Salzburger Lese-Screening für die Klassenstufen 1-4 (Mayringer & Wimmer, 2005) und bei der Würzburger Leise Leseprobe (Küspert & Schneider, 2002).

Des Weiteren steigt die Lesekompetenz mit zunehmender Jahrgangsstufe an (z. B. Tischler, Daseking & Petermann, 2013; Retelsdorf & Möller, 2008; Lehmann & Lenkeit, 2008; Bos et al., 2004). Diesem Umstand wird unter anderem in jahrgangsspezifischen Normen standardisierter Lesetests Rechnung getragen, beispielsweise beim Salzburger Lese-Screening für die Klassenstufen 1-4 (Mayringer & Wimmer, 2005) und bei Knuspels Leseaufgaben (Marx, 2002).

Daneben kann bezüglich der Lesekompetenz folgendes absteigendes Anspruchsniveau angenommen werden: Gymnasium, Realschule, Hauptschule (z. B. Hohn et al., 2013; Retelsdorf & Möller, 2008). Demgegenüber ist die Lesekompetenz bei Schülern aus dem Förderschulbereich häufig vermindert (Hänsel, 2003), hier finden sich oftmals Schüler mit spezifischem Förderbedarf.

Schüler mit spezifischem Förderbedarf

Besonderes Augenmerk bei der Betrachtung von Kindern mit spezifischem Förderbedarf soll auf Schülern der Sprachheilschulen liegen. Sprachheilschulen zeichnen sich durch sprachheilpädagogischen Unterricht aus (Reber & Schönauer-Schneider, 2014; Grohnfeldt & Ritterfeld, 2003), der explizit auf sprachliche Lernprozesse fokussiert (Reber & Schönauer-Schneider, 2014, S. 325) und bei einer Vielzahl von Auffälligkeiten indiziert ist. Individuelle Förderung kann dabei durch spezielle Unterrichtsgestaltung und außerschulische Maßnahmen erfolgen (Reber, 2014).

So fallen beispielsweise Lese- und Rechtschreibstörungen, die nicht am allgemeinen kognitiven Niveau hängen, in den Aufgabenbereich der Sprachheilpädagogik (Osburg, 2003). Die semantisch-lexikalische Störung stellt eine weitere in der Sprachheilbeschulung anzutreffende Auffälligkeit dar. Hierbei bestehen Probleme darin, Äußerungen hinsichtlich der gewählten Wörter zu verstehen (Glück & Elsing, 2014; Glück, 2003). Der verminderte Wortschatz hat negative Auswirkungen auf die Lesekompetenz.

Generell können Sprechstörungen auch auf schriftsprachlicher Ebene Auswirkungen haben, da Lesen eine der Modalitäten ist, in denen sich Sprache äußert (Reber, 2014). Gleichzeitig kann sich der Schrifterwerb positiv auf die Sprachentwicklung auswirken (Reber, 2014). Demzufolge kann bei der Schülerschaft der auf Sprechstörungen spezialisierten Sprachheilschulen von einem spezifischen Förderbedarf ausgegangen werden.

Da auch kognitive Störungen und Intelligenzminderung häufig zusammen mit Sprachentwicklungsstörungen auftreten können (Giel, 2014), muss bei

Untersuchungen in dieser spezifischen Population ein Intelligenzmaß miterhoben werden. Aus den Komorbiditäten mit Störungen aus dem Bereich der allgemeinen kognitiven Fähigkeit folgt jedoch nicht zwangsläufig eine Implikation für die Lernfähigkeit im Sinne der dynamischen Testung (vgl. Hessels, 1997). Es kann vielmehr davon ausgegangen werden, dass die Lernfähigkeit auf dem Niveau der Regelschulen liegt (Guthke & Wiedl, 1996, S. 205).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Im zu konstruierenden dynamischen Lesekompetenztest sollen Kinder mit spezifischem Förderbedarf besondere Berücksichtigung finden. Dabei ist es zur Bestimmung des inkrementellen Nutzens einer dynamischen Testversion gegenüber einer statischen Testversion zielführend, eine Population zu betrachten, bei denen lediglich Förderbedarf hinsichtlich der Lesekompetenz vorliegt, nicht jedoch hinsichtlich der Lernfähigkeit im Sinne des dynamischen Assessments. Es bieten sich demnach insbesondere Sprachheilschüler als Zielpopulation an. Der zu entwickelnde dynamische Lesekompetenztest muss demnach an den Populationen der Regelschüler und Sprachheilschüler validiert werden. Dabei werden die Testwerte des dynamischen Lesekompetenztests mit externen psychologischen Variablen in Verbindung gebracht. Als externe Außenkriterien bieten sich dabei insbesondere Variablen an, die einen Zusammenhang mit der Lesekompetenz aufweisen. Nach den Ausführungen dieses Kapitels sind dies insbesondere das Geschlecht und die Klassenstufe des Schülers sowie seine allgemeinen kognitiven Fähigkeiten, seine Testängstlichkeit und seine allgemeine Ängstlichkeit. Das Vorwissen kann dagegen nicht direkt erfasst und als Außenkriterium in der Validierungsuntersuchung verwendet werden. Stattdessen soll es in der Materialentwicklung besonders berücksichtigt werden.

3.1.3.2 Spezifische Merkmale des Textes

Neben den Merkmalen des Lesers sind auch Merkmale des Textes für den Leseprozess von großer Bedeutung. Einige dieser Merkmale sollen

nachfolgend kurz dargestellt werden: die Textlänge, die Häufigkeit unbekannter Wörter und die Textart.

Textlänge

Obgleich die Länge eines Textes ein wichtiger Einflussfaktor der Textschwierigkeit ist, gibt es relativ wenige empirische Untersuchungen zum Zusammenhang zwischen Textlänge und Textschwierigkeit (Pearson & Hiebert, 2014; Mesmer, Cunningham & Hiebert, 2012). Insbesondere für weniger geübte Leser wie Kinder oder Leser in einer Fremdsprache ist das Lesen längerer Texte ist schwieriger als das Lesen kurzer Texte (z. B. Rasinski & Padak, 2015; Pearson & Hiebert, 2014; Jahani, 2014; Hatcher, 2000; Hiebert, 1999), wobei Redundanzen im Text diesen selbst zwar verlängern, dem Verständnis des Lesers jedoch zuträglich sind (Christmann & Groeben, 1999).

Unbekannte Wörter

Fremde und unbekannte Wörter erhöhen ebenfalls die Schwierigkeit für den Leser (z. B. Pearson & Hiebert, 2014; Schmitt, Jiang & Grabe, 2011; Böhme, 2011; Willenberg, 2007; Christmann & Groeben, 2006). Insbesondere für Leseanfänger ist die Worterkennung für die Leseleistung entscheidend (Bos, Lankes, Prenzel et al., 2003, S. 70f.), sie wird durch viele unbekannte Wörter im Text verringert.

Textart

Lenhard (2013) führt unter anderem die Textgattung als ein bedeutsames Textmerkmal an und unterscheidet zwischen narrativen Texten und Sachtexten. Beide Textgattungen sind als in sich heterogene Gruppen zu verstehen, die auch in Mischformen auftreten können (Belgrad & Pfaff, 2010) und von denen keine der anderen in Hinblick auf leichtes Verstehen überlegen ist (Lenhard, 2013, S. 29). Sachtexte fokussieren auf Informationen, die zu vermitteln und vom Leser zu verarbeiten sind. Die Mehrzahl der in den gymnasialen Lesebüchern der Sekundarstufe I zum Einsatz kommenden Texte sind den Sachtexten zuzuordnen, ihre Verwendungshäufigkeit nimmt von den unteren Klassenstufen bis zu den höheren Klassen der Sekundarstufe I zu (Fischer,

2009). Sie spielen damit im schulischen Kontext eine wichtige Rolle. Insbesondere bei Sachtexten kommt dem bereichsspezifischen Vorwissen eine große Bedeutung zu (Golke et al., 2013; Lenhard, 2013, S. 28).

Narrative Texte hingegen sind häufig mehrdeutiger und von kulturspezifischen Konventionen geprägt (Lenhard, 2013, S. 29). Der narrative Text kann als Alternativentwurf zur Realität gesehen werden, welcher vom Leser gedeutet und in Bezug zur Realität gesetzt wird (Belgrad & Pfaff, 2010; Lenhard, 2013, S. 29). Er erfordert es oftmals auch, „zwischen den Zeilen zu lesen“ (Klicpera et al., 2010). Dem Leser wird nicht alles explizit mitgeteilt, er muss eigenständige Schlüsse ziehen, also Inferenzen bilden (Klicpera et al., 2010, S. 73). Tendenziell scheint damit dem Ziehen von Inferenzen bei narrativen Texten eine größere Bedeutung zuzukommen, während das Vorwissen bei Sachtexten in der Tendenz eine größere Rolle einnimmt.

Insbesondere ist die Unterscheidung zwischen diesen Textarten in Hinblick auf die Gültigkeit des Textes für den Leser von Relevanz. Bei Sachtexten steht die Vertrauenswürdigkeit im Vordergrund, es geht um die Frage, inwieweit der Text glaubwürdig ist. Dagegen ist dem Leser bei Texten über fiktionale Figuren bewusst, dass hier nicht die faktuale Welt, sondern eine imaginierte Fiktion im Text abgebildet wird. Beim Lesen wird jedoch so getan, als ob die imaginierten Figuren des Textes und ihre Handlungen in der Realität spielen (Belgrad & Pfaff, 2010). Trotz dieser Unterschiede ist davon auszugehen, dass die Struktur des Lesens für beide Textsorten einheitlich ist (Christmann & Groeben, 1999, S. 147). Lediglich bei besonderen Gruppen von Lesern kann angenommen werden, dass die Lesekompetenz maßgeblich von der Textart determiniert wird, beispielsweise bei Lesern, die im autistischen Spektrum zu verorten sind und für die Sachtexte oftmals leichter zugänglich sind als viele narrative Texte (Bodner, Engelhardt, Minshew & Williams, 2015; Brown, Oram-Cardy & Johnson, 2013).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Für den zu entwickelnden Lesekompetenztest sind sowohl Sachtexte als auch narrative Texte von Relevanz. Beide Textarten sollen im Test berücksichtigt

werden. Auch wenn dabei keine Unterschiede in Bezug auf die Leseleistung zu erwarten sind, so müssen die textartspezifischen Besonderheiten bei der Konstruktion des Testmaterials angemessen berücksichtigt werden. Kenntnisse kulturspezifischer Konventionen werden oft vom Rezipienten narrativer Texte erwartet und bereichsspezifisches Vorwissen oft vom Leser eines Sachtextes. Sie sind jedoch nicht bei jedem Leser in gleichem Maße vorhanden und können daher in einem Lesekompetenztest zu verzerrten Testwerten führen, wenn sie Gegenstand der zu erbringenden Testleistung werden. Ihnen muss daher bei der Aufgabenkonstruktion besondere Aufmerksamkeit zukommen.

Daneben soll in Hinblick auf die anvisierte Aufgabenschwierigkeit bei der Konstruktion der Testaufgaben auf der Länge der im Rahmen der Testaufgaben eingesetzten Texte ein besonderes Augenmerk liegen. Die bisher gemachten Ausführungen legen ein ansteigendes Schwierigkeitsniveau bei zunehmender Textlänge nahe. Für weniger geübte Leser wie Schüler aus dem Primarbereich sind daher eher kurze Texte empfehlenswert, die für die Aufgaben zu entwickelten Texte sollen daher möglichst kurz gehalten werden. Dies ist auch im Hinblick auf die anvisierte Population der Kinder mit spezifischem Förderbedarf sinnvoll. Ähnlich positive Auswirkungen sind zu erwarten, wenn Fremdwörter und seltene Wörter im Lesekompetenztest möglichst sparsam verwendet werden.

Damit lassen sich aus den textspezifischen Einflussfaktoren auf die Lesekompetenz erste formale Anforderungen an die Lesetestaufgaben festlegen. Ihre inhaltliche Ausrichtung muss dagegen noch genauer eingegrenzt und hergeleitet werden. Dazu müssen in einem nächsten Schritt die beim Lesen ablaufenden Prozesse nochmals genauer betrachtet werden.

3.2 Inferenzen

Nachdem die Lesekompetenz bislang so dargestellt wurde, dass die inhaltliche Breite des Konstrukts zur Geltung kam, soll nachfolgend den auf der Textebene ablaufenden Prozessen ein besonderes Augenmerk zukommen. Diese sind im Sinne der Lesekompetenzdefinition (vgl. Kapitel 3.1) von großer Relevanz, da sie dem Verständnisaspekt im Besonderen Rechnung tragen, indem sie direkt Bezug nehmen auf die „Fähigkeit, geschriebene Texte […] zu verstehen“ (Artelt, Baumert et al., 2001). Unter den Verarbeitungsschritten auf der Ebene der Texte ist hierbei die Inferenzbildung besonders wichtig, insbesondere bei der Bildung lokaler Kohärenzen.

Die Fähigkeit, Inferenzen zu bilden gilt als ein wichtiger Prädiktor der Lesekompetenz (z. B. Oakhill & Cain, 2004). Sie ist insbesondere bei narrativen Texten wichtig, in denen häufig nicht alle Informationen explizit im Text genannt werden (Klicpera et al., 2010, S. 73). Inferenzbildung kann damit als „ein Vorgang des deduktiven Denkens oder Schlussfolgerns“ (Klicpera et al., 2010, S. 73) begriffen werden und ist somit als hierarchiehoch im Sinne von der in Kapitel 3.1.2 beschriebenen Einteilung von Richter und Christmann (2002) zu verstehen. Der in Kapitel 3.1.3.1 aufgekommene Gedanke der „Verarbeitung verbal kodierter Problemstellungen“ (Rost & Schilling, 2006, S. 452; Rost, 1987) findet sich damit in der Bildung von Inferenzen wieder. Mit Rückgriff auf die in Kapitel 3.1.2 gemachten Angaben sind allgemeine kognitive Fähigkeiten für die Inferenzbildung damit von besonderer Bedeutung und ihre Berücksichtigung im zu konstruierenden Lesekompetenztest mit Blick auf die Definition der Lesekompetenz (Kapitel 3.1.1) gerechtfertigt. Mit Hilfe von Inferenzen ist es dem Leser also möglich, ein Verständnis des zu rezipierenden Textes zu erlangen. Sein Vorwissen spielt dabei eine wichtige Rolle (McNamara, 2001; McNamara, Miller & Bransford, 1996).

Über die konkrete Anzahl der Inferenzen und ihre Taxonomie herrscht keine Einigkeit (Chikalanga, 1992). So nennen beispielsweise Graesser, Singer und Trabasso (1994) 13 Klassen von Inferenzen in narrativen Texten: *referential, case structure role assignment, causal antecedent, superordinate goal,*

thematic, *character emotional reaction*, *causal consequence*, *instantiation of noun category*, *instrument*, *subordinate goal-action*, *state*, *emotion of reader* und *author's intent* (Graesser et al., 1994). McNamara et al. (1996) berichten dagegen auch andere Arten von Inferenzen, wie *atypic case-filling inferences*, *instantiation inferences* und *nonsalient property inferences*. Klicpera et al. (2010) unterscheiden zwischen notwendigen Inferenzen, die essenziell für das Textverständnis sind und weiterführenden Inferenzen, die das Textverständnis nur vertiefen und nicht immer gebildet werden müssen. Daneben gibt es noch weitere Unterteilungen der Inferenzen (Chikalanga, 1992), auf die hier jedoch nicht weiter eingegangen werden soll.

Theorien zu Inferenzen

Zwei Theorien zu Inferenzen haben sich als besonders bedeutsam erwiesen und müssen Eingang in den zu entwickelnden Lesekompetenztest finden: die minimalistische Hypothese und die maximalistische Hypothese. Die minimalistische Hypothese postuliert, dass nur logisch zwingende, enge Inferenzen beim Lesen gebildet werden, die für die Bildung lokaler Kohärenzen benötigt werden. Dies geschieht weitgehend automatisch (McKoon & Ratcliff, 1992). Die maximalistische Inferenztheorie postuliert hingegen, dass nur sehr weite Schlussfolgerungen gezogen werden (Graesser et al., 1994). Die empirischen Befunde zu diesen Theorien sind nicht einheitlich, jedoch werden insbesondere beim strategisch-zielbezogenen Lesen relativ weite (elaborierte) Inferenzen gezogen (Richter & Christmann, 2002).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

In Hinblick auf die Bedeutung und theoretische Verankerung der Inferenzbildung kann eine Berücksichtigung der Inferenzen im zu konstruierenden Lesekompetenztest als legitimiert angesehen werden. Sowohl die elaborierten Inferenzen als auch die automatisch ablaufenden Inferenzen haben nach der minimalistischen und der maximalistischen Hypothese ihre Berechtigung und sollen in Aufgaben des zu konstruierenden dynamischen Lesekompetenztest Eingang finden.

Aus der Vielzahl verschiedenster Inferenzen folgt für die konkrete Testentwicklung, dass nicht alle Inferenzen gleichermaßen berücksichtigt werden können. Vielmehr muss eine Auswahl getroffen werden, die neben der zielgruppenspezifischen Passung, beispielsweise hinsichtlich ihrer Komplexität, als theoretisch und empirisch abgesichert angesehen werden kann. Zu diesem Zweck soll nachfolgend ein theoretisches Fundament aufgebaut werden, auf dessen Basis die Inhalte des zu konstruierenden Lesekompetenztests spezifiziert werden können. In den nächsten Kapiteln werden daher zwei umfassende theoretische Modelle dargestellt, in welchen das Verständnis des zu rezipierenden Textes kognitionspsychologisch betrachtet wird: das *construction-integration model* und das *event-indexing model*. Anhand dieser Theorien sollen die Inhalte des dynamischen Lesekompetenztests genauer spezifiziert werden.

3.3 *Construction-integration model*

Nach dem *construction-integration model* von Kintsch (1988, 1998) sind zwei Schritte für den Leser essentiell, um Texte zu verstehen: Zunächst werden aus dem Input des Textes mentale Repräsentationen aufgebaut (*construction*) und diese dann zu einem kohärenten Ganzen verbunden (*integration*) (Kintsch, 1988). Viele empirische Studien sprechen für die Evidenz des Modells (z. B. Mason & Just, 2004; Caillies, Denhière & Kintsch, 2002; Schmalhofer, McDaniel & Keefe, 2002), welches nachfolgend genauer dargestellt werden soll.

Nach dem *construction-integration model* können drei Ebenen der Repräsentation unterschieden werden: die Oberflächenrepräsentation, die propositionale Repräsentation und das Situationsmodell (Eysenck & Keane, 2006, S. 388-389). Diese interagieren miteinander (Schmalhofer & Glavanov, 1986) und sollen nachfolgend kurz erläutert werden.

Oberflächenrepräsentation

Die Oberflächenrepräsentation bildet den genauen Wortlaut und die syntaktische Struktur eines Textes ab (Eysenck & Keane, 2006). Sie wird durch stark automatisierte syntaktische und lexikalische Prozesse gebildet (van Dijk & Kintsch, 1983). Für das Verständnis eines Textes ist sie von untergeordneter Bedeutung (Christmann, 2006), sie ermöglicht es aber, Textteile wortwörtlich zu wiederholen (Schnotz, 2006) und ist Voraussetzung für die Bildung weiterer Repräsentationen.

Propositionale Repräsentation

Die propositionale Repräsentation wird auch Textbasis genannt (Eysenck & Keane, 2006). Sie baut auf der Oberflächenrepräsentation auf und zielt auf den semantischen Gehalt des Gelesenen ab. Dabei wird der Sinn des Gelesenen in Propositionen zerlegt (Eysenck & Keane, 2006). Beim Lesen eines Textes werden sukzessive immer mehr Repräsentationen aufgebaut, zum einen werden sie direkt aus dem Text übernommen, zum anderen werden sie über Brückeninferenzen gebildet (Eysenck & Keane, 2006, S. 387).

Brückeninferenzen sind kohärenzstiftend, da sie als „Brücken“ zwischen Repräsentationen zu verstehen sind. Sie verknüpfen einen anaphorischen Ausdruck mit einem bereits eingeführten Ausdruck, etwa durch ein Pronomen (Artelt et al., 2005, S. 17). Die Verbindung zwischen Oberflächen- und Tiefenstruktur wird mittels Transformationsregeln hergestellt, die angeben, wie die Tiefenstruktur oberflächenstrukturell realisiert werden kann (Christmann, 2006).

Beispielsweise kann der Satz „*Der Hund beißt den Briefträger.*“ nach dem propositionalen Modell von (Kintsch, 1998) folgendermaßen dargestellt werden:

BEISSEN (Hund; Briefträger);

Dieselbe Repräsentation erhält man aber auch bei dem Satz „*Der Briefträger wird vom Hund gebissen.*“, der eine andere Oberflächenstruktur aufweist. Trotz ihrer unterschiedlichen Oberfläche ist die propositionale Bedeutung, also das den Sätzen zu Grunde liegende Konzept gleich.

Die einzelnen Propositionen und ihre unmittelbaren Verknüpfungen bilden die Mikrostruktur eines Textes. Aus ihr geht die Makrostruktur des Textes hervor, die die Mikrorepräsentationen auf einer abstrakteren Ebene zusammenfasst (Kintsch, 1992). Dies geschieht an Hand von drei Regeln: Löschen von Propositionen, die irrelevant für spätere Propositionen sind (*deletion*), Ersetzen mehrere Propositionen durch eine allgemeinere Proposition (*generalisation*) und Ersetzen mehrere Propositionen durch eine einzige Proposition, die eine notwendige Konsequenz dieser Propositionen darstellt (*construction*) (Kintsch, 1998). Durch diese koheränte Verknüpfung ist die semantische Repräsentation für den Leser ressourcenschonender. Christmann (2006) weist darauf hin, dass diese drei Regeln auch rekursiv auf bereits gebildeten Makropropositionen angewendet werden, um die Repräsentation der Textbedeutung weiter zu verdichten. Das Vor- und Weltwissen des Lesers spielt bei der Bildung dieser Makrostrukturen eine wichtige Rolle (Christmann, 2006, S. 615).

Mentales Situationsmodell

Das mentale Situationsmodell ist eine ganzheitliche Beschreibung dessen, was auf propositionaler Ebene repräsentiert wird (vgl. Schnotz, 2006, S. 228). Aus den mentalen Repräsentationen wird ein mentales Modell geformt, das als einheitliches „Bild im Kopf“ die im Text beschriebene Situation abbildet (Schnotz, 2006, S. 227). Das Situationsmodell basiert auf den propositionalen Repräsentationen. Propositionale Repräsentationen enthalten alle relevanten Entitäten eines Satzes, gleichgültig ob diese vorhanden sind oder nicht. Dies unterscheidet sie von Situationsmodellen, bei denen nicht vorhandene Entitäten nicht berücksichtigt werden (Eysenck & Keane, 2006, S. 391). Zwar gibt es keine Beweise für die Existenz mentaler Modelle, jedoch lassen sich viele empirische Beobachtungen mit der Annahme ihrer Existenz leichter erklären (Schnotz, 2006, S. 228).

Beim Bilden der mentalen Repräsentationen der im Text beschriebenen Situationen legt der Leser je nach Zielsetzung und zu rezipierenden Text den Fokus auf verschiedene Aspekte des Modells. Hierbei spielt bereits vorhandenes Wissen des Lesers eine wichtige Rolle. Erinnerungen an vergangene Situationen aus dem episodischen Gedächtnis (van Dijk & Kintsch, 1983) werden ebenso genutzt wie deklaratives Weltwissen (Noordman & Vonk, 1998) und Sprachwissen (Kintsch, 1998, S. 103). Daneben fließen auch Skripte und Schemata des Lesers in das mentale Modell mit ein (McNamara et al., 1996). So werden beispielsweise räumliche Informationen des Texts verarbeitet und mit nicht-räumlichen Informationen in Zusammenhang gebracht. Eine weitere wichtige Dimension ist die Kausalität (Noordman & Vonk, 1998), da sie für die Kohärenzbildung eines Situationsmodells von zentraler Relevanz ist (van den Broek & Gustafson, 1999).

Die bei der Bildung eines Situationsmodells ablaufenden Prozesse und das daraus entstehende „Bild im Kopf“ sind damit hochgradig vom einzelnen Individuum abhängig. Damit erklärt erst diese Textrepräsentationsebene, warum ein und derselbe Text bei unterschiedlichen Personen zu unterschiedlichen mentalen Repräsentationen und damit auch zu unterschiedlichen Interpretationen führen kann (vgl. Schmidthals, 2005, S. 45).

Situationsmodelle weisen nach Rinck (2000) zwei Besonderheiten auf: Erstens sind sie spezifische Repräsentationen einer Situation und ihrer Veränderungen, was sie von generelleren psychologischen Konstrukten wie Schemata und Skripts unterscheidet. Zweitens sind sie flexibel und zielabhängig: Art und Auflösungsgrad der repräsentierten Information können sehr unterschiedlich sein. Die hierfür verantwortlichen Faktoren sind bisher noch kaum genauer untersucht.

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Das *construction-integration-model* beschreibt drei Arten von Repräsentationen, die vom Leser gebildet werden. Ähnlich der Definition der Lesekompetenz wird auch beim *construction-integration-model* eine hohe Komplexität deutlich. Für ein auf das verstehende Lesen abzielende Testkonstrukt, welches auf hierarchiehohen Prozessen auf Textebene basiert, muss nach dieser Theorie ein kohärentes Situationsmodell des gelesen Textes gebildet werden. In ihm sind alle bisherigen Schritte, die für das verstehende Lesen notwendig sind, vereint. Es würde sich daher anbieten, einen Indikator in den zu entwickelnden Lesekompetenztest aufzunehmen, der auf das Situationsmodell abzielt. Indikatoren, die nur auf die Prozesse auf Buchstaben-, Wort- oder Satzebene abzielen, sind nicht als ausreichend anzusehen.

Auch das *construction-integration-model* berücksichtigt die Rolle der Inferenzen beim Lesen, was für eine Umsetzung von Inferenzen im zu konstruierenden Lesekompetenztest spricht. Brückeninferenzen gelten hierbei als besonders zentral für den Aufbau einer propositionalen Repräsentation. Dies erinnert an die in Kapitel 3.1.2 beschriebenen Prozesse auf Satzebene, wo identifizierte Wörter in kohärente Sinnstrukturen überführt werden. Brückeninferenzen sind damit fundamental, um den Sinn des Gelesen zu verstehen. Im zu entwickelnden Lesekompetenztest soll diesem Umstand dadurch Rechnung getragen werden, dass gezielt Informationen abgeprüft werden, die darauf abzielen, Sätze so miteinander zu verbinden, dass eine Sinnstruktur aufgebaut wird.

Eine weitere Informationsart, deren Umsetzung sich im zu konstruierenden Test anbietet, ist die die Kausalität. Sie ist zentral für die Kohärenzbildung eines Situationsmodells. Daneben scheinen auch andere Informationen beim Aufbau eines mentalen Situationsmodells berücksichtigt werden zu müssen, beispielsweise räumliche Informationen. Nachfolgend soll theoriegeleitet begründet werden, welche konkreten Informationen im zu entwickelnden Lesekompetenztest erfragt werden sollen. Das theoretische Fundament hierzu ist Gegenstand des nächsten Kapitels.

3.4 *Event-indexing model*

Ein in Zusammenhang mit Inferenzen bedeutsames kognitionspsychologisches Modell ist das *event-indexing model,* welches sich nur auf narrative Texte bezieht. Auch nach diesem Modell werden vom Leser Situationsmodelle generiert, in dem einzelne *events* miteinander in Beziehung gesetzt werden (Zwaan, Langston & Graesser, 1995).

Die so entstehenden Situationsmodelle können als hochdimensional angesehen werden, wobei insbesondere fünf Dimensionen zu berücksichtigen sind, nach denen die *events* zu einem Situationsmodell zusammengefügt werden (Zwaan, Langston et al., 1995): Raum, Zeit, Kausalität, Intentionen und Protagonist (Zwaan & Radvansky, 1998). Experimentalpsychologische Befunde stützen die Annahme, dass ein Situationsmodell räumliche (z B. Rinck, Williams, Bower & Becker, 1996; Zwaan & van Oostendorp, 1993), zeitliche (z. B. Zwaan, 1996; Bestgen & Vonk, 1995) und kausale (z. B. van den Broek & Lorch, 1993; Zwaan, Magliano & Graesser, 1995) Komponenten enthält und den Protagonisten (z. B. Hakala & O'Brien, 1995) und dessen Intentionen (z. B. Dopkins, 1996; Suh & Trabasso, 1993) berücksichtigt.

Der Leser überwacht beim Lesen diese Dimensionen gleichzeitig (Zwaan, Magliano et al., 1995). Wenn es in einer dieser Dimensionen zu einer Veränderung kommt, also eine Diskontinuität auftritt, dann muss das Situationsmodell aktualisiert werden (Eysenck & Keane, 2006, S. 392). Die Diskontinuitäten werden von Graesser, Millis und Zwaan (1997) wie folgt beschrieben:

1. Räumlich: Die neue Situation findet an einem anderen Ort statt.

2. Zeitlich: Die neue Situation liegt weiter in der Zukunft als die aktuelle Situation oder fand vor der aktuellen Situation statt.

3. Kausal: Es gibt keinen kausalen Zusammenhang zwischen der aktuellen und der neuen Situation.

4. Intentional: Die neue Situation hat einen Bezug zu den Zielen der Figur, unterscheidet sich jedoch von der aktuellen Situation.

5. Protagonistisch: Die neue Situation hat einen von der aktuellen Situation abweichenden Protagonisten.

Findet mehr als eine Diskontinuität gleichzeitig statt, so wird dem Leser kognitiv mehr abverlangt, was sich unter anderem in verlängerten Lesezeiten äußert (Rinck & Weber, 2003).

Inwieweit diese Dimensionen miteinander interagieren und welche Auswirkungen dieses Zusammenspiel auf das Verständnis des Lesers hat, ist jedoch bislang noch nicht hinreichend geklärt. Die empirischen Befunde hierzu sind widersprüchlich (Therriault & Rinck, 2007; Eysenck & Keane, 2006). So wäre beispielsweise zu erwarten, dass sich die durchschnittlichen Lesezeiten insbesondere dann verlängern, wenn der Text derart verändert wird, dass er nicht mehr eine sondern zwei Diskontinuitäten erhält. Jedoch führte die Aufnahme einer zusätzlichen Diskontinuität in einer empirischen Studie nicht automatisch zu verlängerten Lesezeiten (Eysenck & Keane, 2006; Rinck & Weber, 2003).

Die theoretische Annahme, dass diese Dimensionen unabhängig voneinander überwacht werden (Eysenck & Keane, 2006, S. 392), wird von empirischen Befunden in Frage gestellt (Rinck & Weber, 2003). Therriault und Rinck (2007) weisen darüber hinaus darauf hin, dass die Dimensionen Kausalität und Intentionen als Dimensionen zweiter Ordnung betrachtet werden können, da sie von den Dimensionen erster Ordnung nicht unabhängig sind. So ist beispielsweise in der Kausalität eine zeitliche Struktur feststellbar: auf eine Ursache (U) folgt eine durch U bedingte Auswirkung (A). Je kürzer der Zeitabstand zwischen U und A, umso offensichtlicher ist der kausale Zusammenhang (Therriault & Rinck, 2007). Die Einführung von Dimensionen zweiter Ordnung würde bedeuten, dass bei der Überwachung von Kausalität eine zeitliche Dimension stets mit überwacht wird, eine Annahme, die in einem

Spannungsverhältnis zur These steht, dass die Dimensionen unabhängig voneinander überwacht werden.

Bei der Überwachung der einzelnen Dimensionen steht die räumliche Dimension eher im Hintergrund (Zwaan, Radvansky, Hilliard & Curiel, 1998). Es kann dagegen empirisch abgesichert werden, dass neben der Überwachung des Protagonisten auch die zeitliche Dimension eine besondere Relevanz aufweist (Therriault & Rinck, 2007). Selbst wenn die Leser instruiert werden, andere Aspekte zu beachten, werden diese beiden Dimensionen von ihnen überwacht - ein Effekt, der bei der räumlichen Dimension in dieser Form nicht zu finden ist (Therriault, Rinck & Zwaan, 2006).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Für die Bildung von Situationsmodellen sind nach dem *event-indexing-model* verschiedene Dimensionen von theoretischem Interesse, dazu gehören Raum, Zeit und Kausalität. Diese drei Dimensionen finden auch ihren Eingang in empirischen Studien (z. B. Zwaan, Langston & Graesser, 1995; Graesser, Singer & Trabasso, 1994). Sie sind überdies nicht von einer (protagonistischen) Figur des Textes abhängig und können auch in Texten gefunden werden, in denen der Protagonist oder die Intention keine Rolle spielen. Damit wirken sie der Einschränkung entgegen, dass das *event-indexing-model* explizit nur auf narrative Texte abzielt. Sie sind damit potentiell als Inhalte des zu konstruierenden Lesekompetenztests geeignet. In welcher Art und Weise die Inhalte im Test erfragt werden sollen, muss jedoch noch spezifiziert werden.

Insgesamt kann darüber hinaus für das Ziehen von Inferenzen nicht davon ausgegangen werden, dass sich einzelne Aufgaben, die auf spezifische Inferenzarten (temporale Inferenz, kausale Inferenz,...) abzielen, trennscharf voneinander abgrenzen lassen. So können beispielsweise Aufgaben, bei denen kausale Inferenzen gezogen werden nicht als eindimensional angesehen werden, da sie auch die Überwachung zeitlicher Aspekte umfassen. Innerhalb einer Aufgabenart soll jedoch angestrebt werden, eine angemessene Homogenität der Testaufgaben zu erreichen.

3.5 Messung der Lesekompetenz

Neben den bisher aufgeführten theoretischen Überlegungen soll nachfolgend auch der empirische Zugang zur Messung der Lesekompetenz betrachtet werden. Die Erfassung der Lesekompetenz zu diagnostischen Zwecken erfolgt in der Praxis oftmals an Hand sogenannter Kompetenzstufen (Lenhard, 2013). Mit diesen lässt sich beschreiben, „welche Anforderungen Schülerinnen und Schüler mit einem bestimmten Testwert mit einiger Sicherheit bewältigen können“ (Artelt, Schneider & Schiefele, 2002).

Dieser Ansatz findet unter anderem Anwendung in den Schulleistungsstudien PISA (z. B. Drechsel & Artelt, 2007) und IGLU (z. B. Bos, Lankes, Prenzel et al., 2003). Den theoretischen Rahmen der Kompetenzstufenmodelle spannt die probabilistische Testtheorie auf (Bos, Lankes, Schwippert et al., 2003, S. 87). Bei diesem Ansatz wird die Fähigkeitsskala in mehrere Abschnitte, die Kompetenzstufen, unterteilt. Die Aufgaben der einzelnen Abschnitte unterscheiden sich zum einen hinsichtlich ihrer Schwierigkeit, d. h., die Wahrscheinlichkeit, diese Aufgaben bei einem fest definierten Fähigkeitsniveau richtig zu lösen, ist unterschiedlich groß. Daneben zielen die Aufgaben der einzelnen Abschnitte auf unterschiedliche Anforderungen ab, die an den Testanden gestellt werden. Kann ein Testand eine hohe Anforderung bewältigen, dann erreicht er eine hohe Kompetenzstufe und es ist wahrscheinlich, dass er auch Aufgaben bewältigen kann, die aus einer niedrigeren Kompetenzstufe stammen.

Eine tabellarische Übersicht über die verschiedenen Stufenmodelle der Lesekompetenz findet sich auf Seite 8 in Lehmann, Peek und Poerschke (2006). Nachfolgende Tabelle stellt einen für diese Arbeit relevanten Ausschnitt dieser Übersicht dar. Hierbei werden nicht alle Kompetenzstufen aller Modelle dargestellt. Wie ihr entnommen werden kann, hat sich das Konzept der Kompetenzstufen im Bereich des Lesens in unterschiedlichsten Altersklassen bewährt (vgl. Baumert et al., 2001; Lehmann, Peek & von Stritzky, 1995; Elley, 1994; OECD & Statistics Canada, 1995).

Tabelle 2: Stufenmodelle des Leseverständnisses nach Lehmann et. al. (2006)

HAMLET 3-4 nach Lehmann et al. (1997)	**IGLU nach Bos et al. (2003)**	**Hamburger Lesestudie nach Lehmann et al. (1995)**	**IEA Reading Literacy Study nach Elley (1994)**	**PISA nach Baumert et al. (2001)**	**IALS nach OECD/Statistics Canada (1995)**
1. *Einfache Informationen auffinden*	1. *Gesuchte Wörter in einem Text erkennen*	1. *Elementares Leseverständnis* Fähigkeit, einem Text einfache Informationen zu entnehmen	1. *Verbatim Match / Locating Information* Wiederfinden einer identischen wortwörtlichen Formulierung	1. Ausdrücklich angegebene Informationen lokalisieren…	1. *Locating* (Auf-)finden; Identifizieren
3. *Kombinieren / Rekonstruieren*	3. *Implizit im Text enthaltene Sachverhalte auf Grund des Kontextes erschließen*	2. *Generalisiertes Leseverständnis* Fähigkeit, die in einem Text enthaltenen Informationen unabhängig von der konkreten sprachlichen Formulierung zu entnehmen	2. *Paraphrase / Following Instructions* Wiederfinden einer Information, die im Text anders formuliert ist	3. Einzelinformationen und dabei z. T. auch die Beziehungen dieser Einzelinformationen untereinander beachten, die mehrere Voraussetzungen erfüllen…	3. *Integrating* Zusammenfassen; in Beziehung setzen
4. *Verknüpfen / Schluss-folgern*	4. *Mehrere Textpassagen sinnvoll miteinander in Beziehung setzen*	3. *Evaluatives Leseverständnis* Fähigkeit, aufbauend auf der Gesamtinformation eines Textes eigenständige Schlussfolgerungen und Interpretationen zu liefern	4. *Inference* Schlussfolgern, interpretieren, Zusammen-hänge herstellen	5. Verschiedene tief eingebettete Informationen lokalisieren und geordnet wiedergeben	4. *Generating* Interpretieren

Müller und Richter (2013) ziehen Parallelen zwischen den Kompetenzstufen des Lesens und den Repräsentationsebenen nach dem Modell von Kintsch. Bei Aufgaben mit niedriger Kompetenzstufe sind Oberflächenrepräsentation und Textbasis als benötigte Repräsentationsebenen vollkommen ausreichend. Bei Aufgaben höherer Kompetenzstufen, bei denen auf implizit im Text enthaltene Informationen abgezielt wird, müssen die vorhandenen Informationen wissensgestützt interpretiert und kombiniert werden, wofür der Leser mindestens eine propositionale Repräsentation, häufig aber auch ein Situationsmodell benötigt.

Ein Vorteil der Kompetenzstufenmodellierung ist laut Lenhard (2013), dass die einzelnen Testanden nicht mit ihrer Bezugsgruppe verglichen werden. Statt einer relativen Einordnung des Testanden an einer Norm kann mit dem Ansatz der Kompetenzstufen erfasst werden, ob der Testand ein bestimmtes Ziel oder eine bestimmte Fähigkeit erreicht oder nicht. Dies erleichtert die Interpretation und Kommunizierbarkeit der jeweiligen Befunde (vgl. Lenhard, 2013, S. 76).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Die in Tabelle 2 dargestellten Kompetenzstufen sind als empirisch bewährt anzusehen und weisen Parallelen zum *construction-integration model* auf. Durch ihre Unterschiedlichkeit in Hinblick auf die an den Leser gestellten Aufgaben geben sie in gewissem Maße die Heterogenität des Konstrukts der Lesekompetenz wieder. Sie sind daher potentiell als Aufgabenstellungen des zu konstruierenden Lesekompetenztests geeignet. Damit würde der Lesekompetenztest neben dem Bilden von Inferenzen auch auf die Identifikation bestimmter Zielwörter im Text und die Identifikation paraphrasierter Textinformationen abzielen.

Nicht in diesem Stufenmodell berücksichtigt wird jedoch der Inhalt des zu lesenden Textes. So sind beispielsweise nach Kirsch (2001) kausale Informationen als anspruchsvoller für den Leser anzusehen als temporale Informationen, die wiederrum anspruchsvoller als lokale Informationen sind. Daher kann das Kompetenzstufenmodell nicht als alleinige Basis für die

Testkonstruktion ausreichen. Seine Verknüpfung mit anderen Ansätzen erfolgt in Kapitel 5.2.

Mit Blick auf die gelingende Umsetzung des zu entwickelnden dynamischen Lesekompetenztest muss an dieser Stelle noch auf einen Aspekt hingewiesen werden. Gegenstand dieses Kapitels war bislang nur die Messung mittels statischer Verfahren, nicht die Messung mittels dynamischer Verfahren. Die Frage inwieweit eine dynamische Testung der Lesekompetenz überhaupt valide umgesetzt werden kann, wurde bisher noch nicht ausreichend beantwortet. Für die Umsetzung eines dynamischen Lesekompetenztests ist es jedoch zunächst relevant, die Tauglichkeit des Konzepts der Lesekompetenz als Gegenstand der dynamischen Testung zu bestimmen. Eine notwendige Bedingung hierfür ist, dass das Konzept Lesekompetenz intraindividuelle Variabilität aufweist. Eine positive Veränderbarkeit der Lesekompetenz kann als gegeben vorausgesetzt werden, wenn durch gezielte Fördermaßnahmen die Lesekompetenz eines Kindes modifizierbar ist. Dies soll Gegenstand des nächsten Kapitels sein.

3.6 Förderung der Lesekompetenz

Während die kognitionspsychologischen Modelle auf die Beschreibung und theoretische Modellierung von Lesekompetenz abzielen (Kapitel 3.3 und Kapitel 3.4) und die differenziell-psychologische Perspektive ihren Fokus auf die Messung der Lesekompetenz legt (Kapitel 3.5), ist die Förderung derselben ein zentrales Anliegen der pädagogischen Psychologie (Müller & Richter, 2013).

Für eine Vielzahl von Lesekompetenzfacetten wurde mittlerweile eine große Anzahl von Trainingsmaßnahmen und Interventionen entwickelt. Für eine selektive Übersicht sei beispielsweise auf Seuring (2010) verwiesen. In zwei Überblicksarbeiten (Slavin, Lake, Chambers, Cheung & Davis, 2009; Slavin, Cheung, Groff & Lake, 2008) wurden verschiedene Interventionsmaßnahmen bei Kindern und Jugendlichen systematisch auf ihre Effektstärke untersucht. Es zeigte sich, dass die Lesekompetenz mit Hilfe von verschiedensten Trainingsprogrammen verbessert werden kann (Slavin et al., 2009; Slavin et al., 2008). So konnten beispielsweise Trainingsprogramme bei Kindern, die in der ersten Jahrgangsstufe oder einer höheren Jahrgangsstufe waren, im Schnitt das Leseverstehen mit einer Effektstärke von d=0.20 positiv beeinflussen. Für hierarchieniedrigere Lesekompetenzaspekte wurde eine mittlere Effektstärke von d=0.27 erreicht. (Slavin et al., 2009).

Auch das Training der phonologischen Bewusstheit hat einen signifikanten Effekt, nicht nur auf die phonologische Bewusstheit selbst (d=0.81), sondern auch auf die Leseleistung (d=0.51), wie in einer metaanalytischen Untersuchung gezeigt werden konnte (Ehri et al., 2001). Daneben konnte eine metaanalytische Untersuchung auch den Befund liefern, dass das Training der phonologischen Bewusstheit im deutschsprachigen Raum einen signifikanten Einfluss auf spätere Leseleistung hat, es fanden sich kurzfristige (d=0.21) und langfristige (d=0.14) Transfereffekte des Trainings der phonologischen Bewusstheit für den Schriftspracherwerb (Fischer & Pfost, 2015). Die Interventionen im Bereich der Vorläuferfähigkeiten können bereits bei Kindern angesetzt werden, die noch nicht eingeschult worden sind (Blatter, 2014).

Ebenso kann durch gezielte Maßnahmen die Fähigkeit zur Inferenzbildung verbessert werden, wovon insbesondere schwache Leser profitieren (Hall, 2015). Die Fähigkeit zum Ziehen von Inferenzen kann daher ebenfalls als modifizierbar angesehen werden.

Insgesamt kann Lesekompetenz als trainierbar angesehen werden, sie kann durch gezielte Intervention nachweislich positiv beeinflusst werden (Spika, 2015; Rost & Schilling, 2006). Es ergeben sich damit „vielfach fundierte Anhaltspunkte, wie die gezielte, systematische Förderung von Lesekompetenz effektiv […] intensiviert und weiterentwickelt werden kann“ (McElvany & Schneider, 2009, S. 176).

Implikationen in Hinblick auf die angestrebte Testkonstruktion

Aus den Befunden folgt, dass eine intraindividuelle Variabilität der Lesekompetenz angenommen werden kann. Positive Veränderungen der Lesekompetenz können bei Kindern gezielt evoziert werden. Diese Implikation wurde für eine sinnvolle Umsetzung des dynamischen Assessments in Kapitel 2.1 gefordert. Daher kann davon ausgegangen werden, dass ein dynamisches Assessment der Lesekompetenz realisierbar ist. Die hierfür benötigten Grundlagen aus dem Bereich der Leseforschung und ihre Implikation für das zu entwickelnde Testverfahren sollen im Folgenden nochmals zusammengefasst werden.

3.7 Zusammenfassung und Implikationen für die Testkonstruktion

Da die Lesekompetenz eines Kindes durch Interventionen gezielt gefördert und verbessert werden kann (Kapitel 3.6), ist davon auszugehen, dass Veränderungspotential im Sinne der dynamischen Testung gegeben ist und ein dynamischer Lesekompetenztest realisiert werden kann (vgl. Kapitel 2.1). Die zu entwickelnden Testaufgaben zielen dabei konkret auf die Lesekompetenz ab.

Merkmale des Lesers, wie etwa seine kognitiven Fähigkeiten spielen für die Lesekompetenz ebenso eine Rolle wie Merkmale des Textes. Positive Auswirkungen auf die Lesekompetenz haben allgemeine kognitive Fähigkeiten, vorhandenes Vorwissen und weibliches Geschlecht. Negative Auswirkungen auf die Lesekompetenz können dagegen bei Testängstlichkeit und allgemeiner Ängstlichkeit angenommen werden. Schüler in niedrigeren Klassenstufen haben gegenüber Schülern aus höheren Klassenstufen eine verminderte durchschnittliche Lesekompetenz (Kapitel 3.1.3.1). Diese Variablen können im Rahmen der Testvalidierung dafür eingesetzt werden, die Validität des zu entwickelnden dynamischen Lesekompetenztest zu überprüfen. Ein Test der Lesekompetenz sollte Zusammenhänge mit diesen Variablen aufzeigen, die den hier berichteten Befunden entsprechen.

Da die Lesekompetenz von Merkmalen des Lesers beeinflusst wird, lassen sich spezifische Personengruppen identifizieren, bei denen eine Förderung der Lesekompetenz angezeigt ist. Oftmals finden sich Kinder mit spezifischem Förderbedarf im Förderschulbereich. Von besonderem Interesse ist hierbei die Sprachheilbeschulung. Obgleich bei Sprachheilschülern nicht generell von einer vorliegenden Lernbehinderung gesprochen werden kann und auch die Lernfähigkeit im Sinne des dynamischen Assessments nicht vermindert ist, kann bei Sprachheilschülern von einer tendenziell verminderten Lesekompetenz gegenüber Schülern der Regelschule ausgegangen werden (Kapitel 3.1.3.1). Sie sind demnach neben den Grundschülern eine Zielpopulation, für die das zu konstruierende Testverfahren validiert werden soll (vgl. Kapitel 8.2 und Kapitel 8.4).

Insbesondere in Hinblick auf die Population mit spezifischem Förderbedarf ist darauf zu achten, dass die im Test verwendeten Texte hinreichend einfach sind. Da mit zunehmender Textlänge und zunehmender Anzahl an unbekannten Wörtern die Schwierigkeit eines Textes für den Rezipienten steigt, können mittels Textlänge und Wortwahl die Schwierigkeitsparameter der zu erstellenden Texte hinreichend kontrolliert werden. Dagegen hat die Textart in der Regel keine direkten Auswirkungen auf die Schwierigkeit des Textes. Es kann davon ausgegangen werden, dass beim Lesen narrativer Texte der Bildung von Inferenzen im Besonderen Bedeutung zukommt. Demgegenüber fordert das Lesen von Sachtexten in besonderem Maße vorhandenes Vorwissen (Kapitel 3.1.3.2).

Lesekompetenz ist als ein komplexes Konstrukt anzusehen, welches sich aus unterschiedlichsten Komponenten zusammensetzt und bei dem unter anderem Prozesse auf Wort-, Satz- und Textebene ineinandergreifen (Kapitel 3.1.2). Diese hohe Dimensionalität der Lesekompetenz kann abgebildet werden durch verschiedene Indikatoren, die unterschiedliche Teilbereiche des Konstrukts der Lesekompetenz erheben. Von besonderer Bedeutung ist hierbei die Fähigkeit, Inferenzen ziehen zu können. Sie kann als durch eine breite empirische Basis und ein theoretisches Fundament abgesichert betrachtet werden (Kapitel 3.2). Brückeninferenzen sind für den Test ebenso zu berücksichtigen wie die Dimensionen Raum, Zeit und Kausalität (Kapitel 3.3 und Kapitel 3.4). Alle Aufgaben, die dieselbe Inferenzart als Gegenstand haben, sollen nach Möglichkeit so konstruiert werden, dass sie als möglichst eindimensional anzusehen sind.

Außer dem Ziehen von Inferenzen sollen noch weitere Aufgabenarten zum Einsatz kommen. Die Basis für die Auswahl der Aufgabenart ist das Kompetenzstufenmodell zur Erfassung der Lesekompetenz, das sich in großen Schulleistungsstudien bewährt hat. Neben dem Ziehen von Inferenzen ist ihm zufolge auch die Identifikation paraphrasierter und nicht paraphrasierter Textinformationen von besonderer Relevanz für den zu entwickelnden Lesekompetenztest (Kapitel 3.5) und soll in den Testaufgaben umgesetzt werden.

Ausgehend von den bisherigen Befunden ist damit das Fundament für die statische Erfassung der Lesekompetenz gelegt. Es wird bei der Entwicklung der Testmaterialien in Kapitel 5 nochmals aufgegriffen. Der auf diesem Fundament zu entwickelnde statische Lesekompetenztest kann im Sinne der in Kapitel 2.1 gemachten Aussage um eine dynamische Komponente erweitert werden. Für eine gelingende Umsetzung dieser Erweiterung muss berücksichtigt werden, inwieweit die dynamische Version des zu entwickelnden Lesekompetenztests sich von der statischen Version unterscheidet.

3.8 Dynamisches Assessments der Lesekompetenz

Dynamische Prinzipien haben sich bislang im Bereich der Intelligenzdiagnostik bewährt (z. B. Krohne & Hock, 2007; Süß, 2005) und sind darüber hinaus auch auf andere Leistungsbereiche übertragbar (Guthke, Beckmann & Wiedl, 2003). Insbesondere können über die beim Lesen ablaufenden kognitiven Prozesse ähnliche Annahmen getroffen werden wie für die Intelligenz (Dörfler, Golke & Artelt, 2010). Analog zu dynamischen Intelligenztests (Guthke & Wiedl, 1996) zeichnen sich dynamische Lesetests dadurch aus, dass sie ein Lesemaß dynamisch erheben. Es soll damit nicht nur die aktuelle Lesekompetenz erhoben, sondern auch die Lernfähigkeit im Kompetenzbereich des Lesens abgebildet werden. Ein zu konstruierender dynamischer Lesekompetenztest muss also beide Konstrukte in angemessener Art und Weise berücksichtigen. Hierfür können die Befunde aus Kapitel 3 als Basis für die Komponente der Lesekompetenz und die Befunde aus Kapitel 2 als Basis für die dynamische Komponente gelten. Sie werden bei der Entwicklung der Materialien (Kapitel 5) miteinander verbunden.

Für die Testkonstruktion ist es hilfreich, die Überlappung beiden Komponenten a priori zu kennen und damit abschätzen zu können, inwieweit beide Konstrukte dasselbe erfassen. Bei starker Korrelation kann beispielsweise davon ausgegangen werden, dass sich beide Konstrukte so ähnlich sind, dass sie mit ein und demselben Testwert erhoben werden können, bei weniger starker Ähnlichkeit bieten sich zwei voneinander getrennte Indikatoren an, deren konkreter Zusammenhang in der Testkonstruktion nochmals besonders berücksichtigt werden muss.

Eine systematische Zusammenschau der Befunde zum Zusammenhang zwischen der Lesekompetenz und der Lernfähigkeit ist bislang noch nicht umgesetzt worden, obgleich erste Studien Hinweise auf Zusammenhänge des dynamischen Assessments mit der statisch erfassten Lesekompetenz geben konnten (z. B. Dörfler, Golke & Artelt, in press; Swanson, 2011; Kantor, Wagner, Torgesen & Rashotte, 2011). Dieser systematische Überblick soll Thema des nächsten Kapitels sein.

4 Metaanalytische Untersuchung zum korrelativen Zusammenhang der statisch erfassten Lesekompetenz mit dem dynamischen Assessment

4.1 Fragestellung

Wie soeben ausgeführt, muss für die gelingende Konstruktion eines dynamischen Lesekompetenztests noch eruiert werden, inwieweit die Lesekompetenz (LK) Überlappungen mit der Lernfähigkeitskomponente des dynamischen Assessments (DA) aufweist. Das primäre Ziel der metaanalytischen Untersuchung ist es, einen systematischen Überblick über die empirischen Arbeiten zu geben, die die Wirksamkeit des dynamischen Testens im Bereich der Lesekompetenz zeigen.

Dabei ist zu erwarten, dass der dynamische Lesekompetenztest stärker mit der LK korreliert als andere dynamische Verfahren, da ein Teil der gemeinsamen Varianz darauf zurückzuführen ist, dass der dynamische Lesekompetenztest Lesen als Gegenstand hat. Er ist damit bereichsspezifisch für das Lesen (bereichsspezifisches dynamisches Assessment, DA_B).

Dynamische Verfahren, die nicht explizit auf den Bereich Lesen abzielen, werden nachfolgend als allgemeines dynamisches Assessment (DA_A) bezeichnet. Ihr Zusammenhang mit der LK wird auch von allgemeinen kognitiven Fähigkeiten bestimmt sein, die im dynamischen Assessment zur Geltung kommen (vgl. Kapitel 2.7 und Kapitel 3.1.3.1). Daraus folgt, dass die hierarchiehohen, auf die Ebene der Texte abzielenden LK-Facetten stärker als die hierarchieniedrigeren LK-Facetten mit DA_A korrelieren sollten, da sie in stärkerem Maße von diesen allgemeinen kognitiven Fähigkeiten abhängen.

Des Weiteren gilt es auch zu berücksichtigen, dass die interessierenden Zusammenhänge durch verschiedene Drittvariablen beeinflusst werden

können. Auf solche gegebenenfalls vorliegenden Moderatoreffekte muss systematisch geprüft werden.

Drei Fragestellungen sind damit im Rahmen der metaanalytischen Untersuchung von besonderer Relevanz:

M.1. Wie hängt die (statisch erfasste) Lesekompetenz (LK) mit der Leistung in einem dynamischen Lesetest (DA_B) und mit der Leistung in einem anderen dynamischen Testverfahren (DA_A) zusammen?

M.2. Welche Facetten der (statisch erfassten) Lesekompetenz LK hängen insbesondere mit der Leistung in einem dynamischen Lesetest (DA_B) und mit der Leistung in einem anderen dynamischen Testverfahren (DA_A) zusammen?

M.3. Welche Variablen moderieren gegebenenfalls diese Zusammenhänge?

Im Rahmen der hier beschriebenen metaanalytischen Untersuchung wird immer von statischer Lesekompetenz (LK) gesprochen, um hervorzuheben, dass es sich bei den dynamischen Lesetests (DA_B) und bei der statisch erfassten Lesekompetenz (LK) um zwei unterschiedliche Konstrukte handelt, die in der Metaanalyse in Zusammenhang gebracht werden sollen. Die statische Lesekompetenz ist damit abzugrenzen von den dynamischen Lesetests, die eine Untergruppe der dynamischen Testverfahren bilden.

4.2 Methodik

Das methodische Vorgehen orientierte sich an Empfehlungen aus der einschlägigen Fachliteratur (Cooper, 2010; Lipsey & Wilson, 2001). Zunächst wurden die zu erfassenden Konstrukte genauer spezifiziert (Kapitel 4.2.1) und diese in einschlägigen Datenbanken gesucht (Kapitel 4.2.2). Aus den gefundenen Studien wurden diejenigen extrahiert, deren Berücksichtigung für die Beantwortung der metaanalytischen Fragestellungen zielführend war. Diese Studien wurden anhand von spezifischen Variablen kodiert (Kapitel 4.2.3). Die so ermittelten Primärstudien wurden zunächst um ihre Artefakte bereinigt, bevor die statistische Überprüfung der Fragestellungen durchgeführt werden konnte (Kapitel 4.2.4).

4.2.1 Spezifikation des Untersuchungsgegenstands

Dynamisches Assessment in der Metaanalyse

In der metaanalytischen Betrachtung sollten ausgehend von der in Kapitel 2 dargestellten Forschungslage nur bestimmte Arten von dynamischem Assessment berücksichtigt werden, um eine ausreichende Homogenität der Primärstudien zu gewährleisten.

So wurde im Hinblick auf hinreichende Quantifizierbarkeit des dynamischen Assessments und seiner empirischen Zusammenhänge der auf Feuerstein beruhende interaktionistische Ansatz für die Metaanalyse nicht berücksichtigt (vgl. Kapitel 2.4.1). Darüber hinaus musste die im dynamischen Test verwendete Rückmeldung als standardisiertes Feedback erfolgen, wobei die Feedbackmodalitäten besondere Berücksichtigung finden sollten.

Das dynamische Assessment (DA) wurde im Rahmen der metaanalytischen Untersuchung nochmals genauer differenziert, es konnte bereichsspezifisch (DA_B) oder allgemein (DA_A) sein. Bereichsspezifisches dynamisches Assessment zielte auf dynamische Lesetests ab, hier war die Lesekompetenz bereits Gegenstand des dynamischen Tests und wurde explizit als solche erhoben. Allgemeines dynamisches Assessment maß dagegen explizit keine

Facette der Lesekompetenz, sondern zielte auf arbeitsgedächtnis-, *reasoning*-, beziehungsweise intelligenznahe Konstrukte ab. Damit wies es eine besondere Nähe zu den allgemeinen kognitiven Fähigkeiten auf und war in Hinblick auf die Bedeutung kognitiver Faktoren für die interessierenden Zusammenhänge im Besonderen zielführend.

Studien, die DA_B erhoben, und Studien, die DA_A erhoben, verwendeten somit beide ein dynamisches Maß. Unabhängig vom konkreten Konstrukt, welches dieses dynamische Maß erfasste (Leseverstehen bei DA_B oder ein allgemeines kognitives Potential bei DA_A) hatten beide Arten von Studien damit eine entscheidende Komponente der dynamischen Testung gemeinsam: die Lernfähigkeitskomponente, die das Alleinstellungsmerkmal des dynamischen Assessments darstellt (vgl. Kapitel 2.1). Es ist damit legitim, sie beide unter dem Begriff dynamisches Assessment (DA) zusammenzufassen.

Dynamisches Assessment aus anderen Bereichen, beispielsweise dynamische Tests aus dem Kompetenzbereich Schreiben, blieben unberücksichtigt. Ihr Mehrwert für die Metaanalyse war begrenzt, da sie weder explizit auf Lesekompetenz abzielen, noch auf Bereiche, die mit allgemeinen kognitiven Fähigkeiten so stark assoziiert waren wie das dynamische Assessment von arbeitsgedächtnis-, *reasoning*- und intelligenznahen Konstrukten.

Lesekompetenz in der Metaanalyse

Die in der Metaanalyse betrachtete Lesekompetenz (LK) wurde pragmatisch definiert und orientierte sich an den Definitionen der jeweiligen Primärstudien. Damit wurde die ganze Bandbreite der Lesekompetenz im Rahmen der metaanalytischen Untersuchung abgedeckt. Die in den Primärstudien gefundenen Lesekompetenzmaße wurden in Anlehnung an der in Kapitel 3.1.2 dargelegten Klassifikation der Lesekompetenzkomponenten (vgl. Tabelle 1) in fünf Klassen unterschiedlicher LK-Facetten gruppiert, die nachfolgender Tabelle entnommen werden können: Texte verstehen, Wortschatz, Dekodierfähigkeit, Vorläuferfähigkeiten und ein Mischscore aus diesen Bereichen. Allen ist gemein, dass sie nicht mittels dynamischer Testung, sondern statisch erfasst wurden.

Tabelle 3: Für die Metaanalyse relevante Facetten der Lesekompetenz

Spezifische Kompetenz	**Kategorisierung nach Richter und Christmann (2002)**	**Beschreibung**
Texte verstehen	hierarchiehoch	Fokus auf Bildung von Inferenzen, verstehendes Lesen
Wortschatz	hierarchieniedrig	Fokus auf Worterkennung, lexikalischer Zugriff nötig
Dekodierfähigkeit	hierarchieniedrig	Fokus auf Nicht-Wörter, Pseudowörter und Buchstaben, kein lexikalischer Zugriff nötig
Vorläuferfähigkeiten	-	Fokus auf phonologische Bewusstheit und andere Vorläuferfähigkeiten
Mischscore aus obigen Bereichen	-	

Die LK-Facette „Texte verstehen“ (LK_T) umfasste in diesem Zusammenhang alle Leseleistungen, bei denen Sinnzusammenhänge und das Ziehen von Inferenzen im Vordergrund standen. Sie war damit die einzige Lesekompetenzfacette, die im Sinne von Richter und Christmann (2002) als hierarchiehoch bezeichnet werden kann. Alle anderen Facetten der Lesekompetenz zielten nicht explizit auf die Textebene ab und sollen nachfolgend mit LK_{NT} bezeichnet werden.

Wortschatz zielte auf das Lesen und Verstehen einzelner Wörter ab. Dies unterschied diese Lesekompetenzfacette von der Dekodierfähigkeit, bei der kein lexikalischer Zugriff benötigt wurde. Vielmehr umfasste die Dekodierfähigkeit die Dekodiergenauigkeit und -geschwindigkeit, welche auch auf der Ebene einzelner Grapheme erhoben werden konnten. Zur Gruppe der Vorläuferfähigkeiten gehörten neben der phonologischen Bewusstheit auch

allgemeine sprachliche Vorläuferfähigkeiten, die bereits an Kinder im Vorschulalter erhoben werden konnten.

Daneben konnten auch Mischscores aus den soeben genannten Bereichen auftreten, etwa wenn ein Summenindex aus einer Messung des korrekten Erkennens von Buchstaben und aus einer Messung zum korrekten Verständnis eines Textes gebildet wurde. Bei der Bildung des Mischscores musste der Tatsache Rechnung getragen werden, dass hierarchieniedrige Fähigkeiten sich auch stets in den hierarchiehöheren Lesekompetenzfacetten wieder finden lassen (vgl. Kapitel 3.1.2). So wird beispielsweise das korrekte Erkennen von Buchstaben und Wörtern immer benötigt, um einen Text zu verarbeiten und seinen Sinn zu durchdringen. In diesem Fall wurde jedoch nicht der Mischscore, sondern die hierarchiehöchste Lesekompetenzfacette als das für die Analyse relevante Lesemaß festgelegt.

4.2.2 Literaturrecherche

Für die Literaturrecherche wurden fünf psychologische Datenbanken (PSYNDEX, PsycINFO, PSYCArticles, PsychSpider, PsyJournals) und neun Dissertationsdatenbanken aus verschiedenen Ländern durchsucht. Tabelle 4 gibt eine Übersicht über die Datenbanken, ihre nationale Zuordnung und ihre Internetadresse. In Tabellen 5 und 6 sind für die psychologischen Datenbanken und die Dissertationsdatenbanken jeweils die Anzahl der Treffer zu finden, die mit bestimmten Schlagworten in der jeweiligen Datenbank ermittelt werden konnten.

Tabelle 4: Übersicht über die verwendeten Dissertationsdatenbanken

Suchmaschine	**Nationale Zuordnung**	**Internetadresse**
OPUS-Metasuche	Deutschland	http://elib.uni-stuttgart.de/opus/gemeinsame_suche.php
DNB	Deutschland	http://www.dnb.de/DE/Wir/Kooperation/dissonline/dissonline_node.html
obv sg	Österreich	http://www.obvsg.at/services/dissertationsdatenbank
Dissexpress	Vereinigte Staaten von Amerika	http://disexpress.umi.com/dxweb
CAUL	Australien	http://adt.caul.edu.au/
DBIS	Vereinigtes Königreich/Irland	http://rzblx10.uni-regensburg.de/dbinfo/detail.php?bib_id=alle&colors=&ocolors=&let=k&tid=0&titel_id=690
Theses Ca	Kanada	http://www.collectionscanada.gc.ca/thesescanada
nz research	Neuseeland	http://aut.ac.nz.libguides.com/content.php?pid=64268&sid=475113
theses fr	Frankreich	http://www.theses.fr/

Bei besonders einschlägigen Studien wurden die jeweiligen Referenzen auf weitere potentiell relevante Literatur geprüft und die jeweiligen (Erst-)Autoren der Studie angeschrieben und nach nicht veröffentlichten Studien gefragt. Ziel dieses Vorgehens war es, weitere unpublizierte Daten zu finden. So sollte eine möglichst umfassende Literaturrecherche gewährleistet sein und dem *Publication Bias* entgegengewirkt werden (Cooper, 2010, S. 246).

Für alle statistischen Analysen und die Interpretation der Ergebnisse wurden die Namen der Primärstudien anonymisiert, um eine verblindete statistische Analyse zu gewährleisten, bei der weder ein berühmter Autorennamen noch eine geachtete Zeitschrift implizit weder in die Bewertung der einzelnen Studien noch in die Interpretation der Ergebnisse miteinfließen konnte. Die Pseudonyme der einzelnen Studien bestehen immer aus einem Buchstaben und einer Zahl zwischen 1 und 100 (z. B. s16). Sind zwei Studien innerhalb derselben Publikation beschrieben, so wird das mit römischen Ziffern kenntlich gemacht (z. B. f27 I, f27 II). Diese Anonymisierung wird auch in dieser Arbeit bei der Darstellung der Befunde beibehalten, eine Aufschlüsselung findet sich in Anhang A.

Tabelle 5: Spezifische Suchbegriffe und Trefferanzahl in psychologischen Datenbanken

Suchbegriffe	**PSYNDEX**	**PsycINFO**	**PSYCArticles**	**PsychSpider**	**PsyJournals**
dynamischer Test + Lesekompetenz	0	0	0	0	0
dyn. Test + Lesefähigkeit	0	0	0	0	0
dyn. Test + Lesefertigkeit	0	0	0	0	0
dynamic test	13	79	3	53	5
dynamic test + reading	7	7	6	12	0
Rückmeldung + Test	208	3	0	2767	137
dynamic assessment + reading	2	74	2	20	2
dyn. Test + Lernfähigkeit	0	0	0	1	1
elaboriertes Feedback	2	0	0	12	1
Feedback + Test	617	7069	511	19337	774
Lesekompetenz + Test	65	1	0	423	58
learning potential assessment	21	328	13	175	23
mediated learning	15	564	13	96	3
testing the limits	49	132	15	159	18
mediated assessment	0	11	0	2	4
assisted learning and transfer	7	28	1	126	4
Suche nach Autoren					
Tzuriel, D(avid)	0	76	1	8	0
Feuerstein, R(euven)	0	47	2	5	0
Resing, W(ilma)	4	62	1	4	1
Campione, J(ospeh)	0	51	9	16	0
Swanson, H(oward)	0	225	19	45	0
Wiedl, K(arl)	0	109	4	37	8
Sternberg, R(obert)	6	769	85	173	7
Lidz, C(arol)	0	48	1	12	0
Ferrara, S(teve)	0	25	0	3	0
Brown	1	134	9	79	0

Tabelle 6: Spezifische Suchbegriffe und Trefferanzahl in internationalen Dissertationsdatenbanken

Suchbegriffe	OPUS-Metasuche	DNB	obv sg	Dissexpress	CAUL	DBIS	Theses Ca	nz research	theses fr
dynamischer Test + Lesekompetenz (dynamic test + reading competence)	0	0	0	0 (3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
dyn. Test + Lesefähigkeit (dyn. Test + reading ability)	0	0	0	0 (13)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
dyn. Test + Lesefertigkeit	0	0	0	0	0	0	0	0	0
dynamic test	29	16	14	40	39	36	23	11	29
dynamic test + reading	0	0	0	40	0	0	0	0	12
dynamic assessment + reading	0	0	0	40	1	1	0	0	2
dyn. Test + Lernfähigkeit (dyn. Test + learning ability)	0	0	0	0 (22)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Lesen + Lernfähigkeit (learning ability + reading)	13	0	233	0 (40)	0 (1)	0(4)	0 (90)	0 (0)	0 (11)
Rückmeldung + Test	702	0	0	0	0	0	0	0	0
Feedback + Test	114	4	0	40	390	525	391	129	1168
elaboriertes Feedback (elaborated feedback)	2	0	0	0 (40)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Lesekompetenz + Test (reading competence + test)	6	1	0	3 (40)	0 (2)	0 (5)	0 (0)	0 (0)	0 (1)
learning potential assessment	0	0	0	40	2	1	0	2	2
mediated learning	2	4	3	40	37	19	20	4	1
testing the limits	2	3	1	40	6	9	8	2	5
mediated assessment	1	0	0	40	1	0	0	0	0
assisted learning and transfer	0	0	0	3	0	1	0	0	0

Anmerkung. Trefferzahlen für die englische Übersetzung der Suchbegriffe sind jeweils in Klammern angegeben.

4.2.3 Literaturauswahl und -kodierung

Literaturauswahl

Anhand der Abstracts und einer ersten Vorselektion ließen sich 88 potentielle Primärstudien identifizieren, die genauer auf ihre methodische Qualität und inhaltliche Passung zur Fragestellung überprüft wurden. Insbesondere ist die Gewährleistung der methodischen Qualität der Primärstudien bei der Durchführung einer Metaanalyse von zentraler Bedeutung. Die Fachliteratur spricht in diesem Zusammenhang vom „*garbage in, garbage out*"-Effekt (Lipsey & Wilson, 2001, S. 9), nach dem die Qualität der Metaanalyse direkt von der Qualität der Primärstudien abhängig ist. So führt auch ein inhaltlich und methodisch unangreifbarer metaanalytischer Ansatz nicht zu aussagekräftigen Ergebnissen, wenn die betrachteten Primärstudien invalide sind (Lipsey & Wilson, 2001). Um diesem Effekt entgegenzuwirken wurden 17 Ausschlusskriterien definiert, die im Folgenden kurz skizziert werden sollen. Eine Übersicht der Kriterien findet sich in Tabelle 7. Durch eine systematische Anwendung dieser Ausschlusskriterien sollte eine hinreichende Homogenität der Primärstudien sichergestellt werden, welche als eine notwendige Voraussetzung für einen sinnvollen Vergleich der Studien miteinander angesehen werden kann (Lipsey & Wilson, 2001, S. 8). Die Selektion wurde von zwei voneinander unabhängigen Ratern durchgeführt. Eine Studie wurde nur dann in die Metaanalyse aufgenommen, wenn sie von beiden Ratern als inhaltlich und methodisch geeignet erachtet wurde. Dies war bei 16 Studien der Fall.

Die problematischen Aspekte, die zu einem Ausschluss der Studie führen können, lassen sich in vier Teilbereiche untergliedern: strukturelle Aspekte der Studie, Erhebung der Zielvariablen, Stichprobenauffälligkeiten und generelle methodische Aspekte. Diese werden nachfolgend genauer dargelegt.

Tabelle 7: Problematische Aspekte potentieller Primärstudien

Strukturelle Aspekte der Studie	**Aspekte bei der Erhebung der Zielvariablen**	**Stichprobenspezifische Aspekte**	**Methodische Aspekte**
Master- oder Bachelorarbeiten	Kein Lesemaß oder Vorläufer der Lesefähigkeit erhoben	Unangemessen hoher Dropout	Boden- oder Deckeneffekte
Aktualität der Studie nicht hinreichend	Kein objektives Leistungsmaß verwendet	Probanden nicht im relevanten Altersabschnitt	Der für die Metaanalyse interessante Zusammenhang wird in Studie nicht deutlich
Artikel nicht auf Deutsch bzw. Englisch publiziert	Dynamisches Assessment nicht feedbackbasiert	Stark eingeschränkte Repräsentativität der Stichprobe	Starke methodische Mängel in Studie
Studie greift auf Stichprobe einer anderen, besser geeigneten Studie zurück	Benotung der in Studie gemessenen Leistung	Stichprobenumfang zu gering für angemessene parametrische Datenauswertung	Berechnung der Effektstärken wegen fehlender Angaben nicht möglich
		Abweichende Muttersprache oder Zweisprachigkeit in Subpopulation	

Zu den strukturellen Aspekten der Studien, die einer Aufnahme in die Metaanalyse entgegen standen, zählte die Art der wissenschaftlichen Arbeit. Bachelor-, Diplom- oder Masterarbeiten wurden nicht berücksichtigt. Untersuchungen vor 1980 wurden ebenfalls nicht berücksichtigt, ebenso wenig Befunde, die weder auf Deutsch noch auf Englisch publiziert wurden.

Ein weiteres Problemfeld der metaanalytischen Forschung sind Abhängigkeiten zwischen den einzelnen Primärstudien. Wenn sich mehrere Publikationen auf denselben Datensatz beziehen, beinhalten sie damit dieselbe Stichprobe und sind nicht mehr voneinander unabhängig. Jede Stichprobe wurde daher nur einmal in Metaanalyse aufgenommen. Welche Studie der Fragestellung der Metaanalyse am meisten entgegen kam, wurde von den beiden voneinander unabhängigen Ratern in gemeinsamer Diskussion entschieden. Damit fanden Studien keinen Eingang in die Metaanalyse, deren Stichprobe bereits in einer anderen, für die Metaanalyse besser geeigneten Studie verwendet wurde.

Die problematischen Aspekte bei der Erhebung der Zielvariablen umfassten vier Bereiche. Erstens wurden Studien nicht berücksichtigt, die kein Lesemaß oder eine Vorläuferfähigkeit der Lesekompetenz statisch erhoben haben. Hierzu gehören unter anderem die Untersuchungen, bei denen Lesen ausschließlich dynamisch, beispielsweise im *train-within-test*-Format erhoben haben. Zweiten schieden alle Studien aus, die kein objektives Leistungsmaß, sondern ein subjektives Leistungsurteil verwendeten, beispielsweise ein subjektives Lehrerurteil. Drittens musste das dynamische Assessment feedbackbasiert sein, um eine hinreichende Nähe zu dem zu konstruierenden feedbackbasierten Testverfahren zu haben. Aus diesem Grund durfte die gemessene Leistung in der Studie auch nicht benotet und damit über den Fokus der Studie hinaus für die Probanden relevant sein.

Mögliche Problemfelder in Hinblick auf die Stichprobenbeschaffenheit fanden ebenfalls ihre Berücksichtigung als Ausschlusskriterien. Studien mit unangemessen hohem Dropout (z. B. 24 % und mehr) wurden von der Metaanalyse ausgeschlossen, ebenso Studien, bei denen die Probanden nicht

im gesuchten Altersabschnitt zwischen etwa 4 und etwa 15 Jahren waren. Damit ist der Alters*range* der Stichproben so gelegt, dass er die Primarstufe und die Sekundarstufe I umfasst, also in etwa den Bereich der Pflichtschulzeit. Studien mit Erwachsenen wurden ausgeschlossen. Dies ist in Einklang mit der von Wygotsky (1978) anvisierten kindlichen Zielpopulation und trägt überdies dem pädagogisch-förderungsdiagnostischen Grundgedanken des Gesamtprojekts im Besonderen Rechnung, da nur Ergebnisse an Kindern und Jugendlichen eine valide Basis für die Entwicklung eines dynamischen Tests für Kinder darstellen. Daneben war eine stark eingeschränkte Repräsentativität der Stichprobe ebenso ein Ausschlusskriterium wie ein Stichprobenumfang, der zu gering für eine angemessene parametrische Datenauswertung war. Von der Mehrheit abweichende Muttersprache(n) oder Zweisprachigkeit in den Stichproben oder in ihren Subpopulationen wurde ebenfalls kontrolliert und diese Information als potentielle Moderatorvariable berücksichtigt. Untersuchungen mit starken diesbezüglichen Auffälligkeiten wurden gegebenenfalls von der Analyse ausgeschlossen.

Allgemeine methodische Problematiken, die zu einer Nichtberücksichtigung der Untersuchungen führten, umfassten starke Boden- oder Deckeneffekte, die entweder direkt berichtet oder auf Grund von besonders extremen Testwerten als sehr naheliegend betrachtet werden konnten. Nicht berücksichtigt wurden außerdem Studien, bei denen der für die Metaanalyse relevante Zusammenhang nicht deutlich wurde (z. B. bei Mediator- oder Moderatoranalysen). Von einer Aufnahme in die Metaanalyse wurde außerdem auch dann abgesehen, wenn starke methodische Mängel in der Studie feststellbar waren (z. B. Konfundierung von Experimental- und Kontrollgruppe, mangelhafte Reliabilität der Messinstrumente) oder die Berechnung der Effektstärken auf Grund von zu vielen fehlenden Angaben nicht möglich war und die fehlenden Informationen vergeblich bei den Autoren der Studien angefordert wurden.

Für die Reduktion der gefundenen Primärstudien waren zwei Rater verantwortlich, die unabhängig voneinander die Studien bewerteten und für jede Studie in Diskussion entschieden, ob sie in der Metaanalyse Berücksichtigung finden sollte. Eine Studie wurde nur dann aus der Analyse

ausgeschlossen, wenn beide Rater der Überzeugung waren, dass diese Studie dem Ziel der Metaanalyse nicht dienlich war. Hierfür wurden die Studien von beiden Ratern nach einem fest vorgegebenen Kriterienkatalog kodiert.

Kodierung und Bewertung der Studien

Im Fokus der Kodierung stand neben der methodischen Qualität der Primärstudien auch der Einfluss potentieller Moderatorvariablen auf die interessierenden Zusammenhänge (Fragestellung M.3.). Daher wurde eine große Anzahl potentieller Moderatoren bei der Kodierung berücksichtigt. Insgesamt wurde nach 60 Kriterien kodiert. Fünf Kriterien erfassten grundlegende Daten, wie die Art der Publikation und fünf Kriterien gaben einen allgemeinen Überblick über die Methodik wie beispielsweise die Art der Untersuchung (Gruppenvergleich, Autoregression,…) und den Untersuchungszeitraum. Auf die Stichprobenbeschaffenheit zielten 13 Kriterien ab, hier wurden beispielsweise das Alter und die Jahrgangsstufe der Probanden erfasst. 23 Kriterien zielten auf die spezifischen Merkmale der Lesekompetenz und des dynamischen Assessments ab, wie die Frage, ob zur Messung der Konstrukte Eigenentwicklungen oder publizierte Testverfahren verwendet wurden oder die Frage nach der genauen LK-Facette. Dem Feedback, welches im Rahmen des DA gegeben wurde, kam eine besondere Bedeutung zu. Ausgehend von Befunden aus der Feedbackforschung (Shute, 2008; Narciss, 2006; Bangert-Drowns, Kulik, Kulik & Morgan, 1991; Kluger & DeNisi, 1996) wurden besonders relevante Feedbackcharakteristiken identifiziert und mit 12 Kriterien erfasst. Hierzu zählte beispielsweise die Frage nach dem Feedbackgeber oder auch in welcher Form das Feedback gegeben wurde. Zwei Kriterien dienen der Zusammenfassung und Bewertung der Studie. Eine vollständige Auflistung aller Kriterien kann bei der Autorin angefordert werden.

Sämtliche Studien wurden von den zwei voneinander unabhängigen Ratern kodiert, vorher übten beide Beurteiler die Ratingprozedur an mehreren Studien ein, um eine adäquate Beurteilerübereinstimmung zu erzielen. Bei abweichenden Beurteilungen durch die Rater wurde durch Diskussion eine Einigung erzielt. Die Interraterreliabilitäten waren für die in der vorliegenden

Studie interessierenden Variablen mit Kappa-Koeffizienten zwischen .618 (Art des dynamischen Assessment) und 1 (restliche Variablen) zufriedenstellend. Der im Vergleich relativ niedrige Kappa-Koeffizient beim Merkmal Art des dynamischen Tests lässt sich teilweise auch durch Mischformen zwischen dem *train-test-train*-Design und dem *train-within-test*-Design erklären (vgl. Kapitel 2.4.2), bei denen die Zuordnung nicht eindeutig war.

4.2.4 Datenaufbereitung und Homogenitätsprüfung

Berechnung der Effektstärken

Insgesamt hatten 15 Studien korrelationsbasierte (*r family*) Effektstärkenmaße. Die Effektstärke der Studie f5 musste jedoch aus dem Regressionsgewicht β berechnet werden, was in diesem Setting unter Umständen problembehaftet sein könnte (Peterson & Brown, 2005). Es wurde daher eine Sensitivitätsanalyse durchgeführt, um zu überprüfen, ob die Aufnahme der Studie mit der β-basierten Effektstärke eine starke Auswirkung auf die aggregierte Effektstärke hatte. Eine qualitative Inspektion der Gesamtdaten mit und ohne f5 widerlegte die Annahme einer starken Auswirkung und die Studie konnte in die Metaanalyse aufgenommen werden.

Korrektur von Artefakten

Dem Ansatz von Hunter und Schmidt (2004) folgend, muss – soweit möglich – um Artefakte korrigiert werden. Jedoch waren nicht alle für eine Artefaktkorrektur benötigten Informationen in den Primärstudien enthalten. Um die noch benötigten Informationen zu erlangen, wurden die jeweiligen Erstautoren der Studien angeschrieben und die jeweils verwendeten Verfahren und ihre Reliabilitäten recherchiert. Für drei Studien (s16, f27 I, f27 II) konnten korrigierte Effektstärken berechnet werden, die nachfolgend verwendet wurden.

Homogenitätsprüfung

Entsprechend den Empfehlungen von Hedges und Vevea (1998) sowie von Hunter und Schmidt (2004) wurden nachfolgend *random-effects models* gerechnet. Das Signifikanzniveau lag einheitlich bei p=.05. Verwendet wurde das von Borenstein, Hedges, Higgins und Rothstein (2015) entwickelte Programm *Comprehensive Meta-Analysis (Version 3).*

Im Rahmen der Homogenitätsprüfung der Punktschätzer der Effektstärken wird, wie von Huedo-Medina, Sánchez-Meca, Marín-Martínez und Botella (2006) vorgeschlagen, neben der Q-Statistik auch der I^2-Index berichtet, welcher Auskunft über das Ausmaß der Heterogenität gibt. Der I^2-Index kann

Werte zwischen 0 und 100 annehmen und gibt den prozentualen Anteil der Streuung in den Punktschätzern der Effektstärken an, welcher auf die Heterogenität der Studien zurückzuführen ist (Higgins & Thompson, 2002). Er ist unabhängig von der konkreten Anzahl der Primärstudien (Higgins, Thompson, Deeks & Altman, 2003).

4.3 Ergebnisse

Tabelle 8 gibt eine Übersicht über die miteinander zu korrelierenden Variablen der 16 Studien, die in der Metaanalyse berücksichtigt werden. Alle Primärstudien haben ein dynamisches Assessment durchgeführt und ein statisches Lesekompetenzmaß erhoben und diese miteinander korreliert. Das durchgeführte dynamische Assessment ist bei acht Studien bereichsspezifisch (DA_B) und bei acht Studien allgemein (DA_A), d.h. nicht explizit auf den Bereich des Lesens abzielend. Fünf Primärstudien messen die auf die Ebene der Texte abzielende, hierarchiehohe Lesekompetenzfacette (LK_T), elf Studien zielen auf andere Lesekompetenzfacetten ab (LK_{NT}).

Tabelle 8: Arten der Lesekompetenz (LK) und des dynamischen Assessments (DA) und ihre Verteilung auf die Primärstudien der Metaanalyse

Art der Lesekompetenz (LK)	**Art des dynamischen Assessments (DA)**	**Anzahl an Studien**
LK_T	DA_A	2
LK_T	DA_B	3
LK_{NT}	DA_A	6
LK_{NT}	DA_B	5

Abbildung 3 zeigt für jede der 16 Primärstudien den jeweiligen Zusammenhang der in der Studie erhobenen LK-Facette mit der in der Studie erhobenen DA-Facette. In jeder Zeile ist eine anonymisierte Primärstudie abgetragen. Die Effektstärken des Zusammenhangs werden aus Gründen der besseren Interpretierbarkeit für jede Studie als Korrelation mit entsprechendem 95 %-Konfidenzintervall berichtet und grafisch dargestellt. Das Aggregat aller 16 Studien ist umrahmt und befindet sich in der untersten Zeile.

Wie aus der letzten Zeile der Abbildung ersichtlich, korreliert DA positiv mit LK. Das entsprechende 95 %-Konfidenzintervall umfasst den Bereich zwischen .269 und .511. Allerdings wurde der Homogenitätstest signifikant

(Q=183.091, df=15, p<.001, I^2=91.807). Daher kann die hier gefundene Gesamtkorrelation nicht interpretiert werden. Stattdessen ist es sinnvoll, die 16 Primärstudien in homogenere Untergruppen zu untergliedern und die Untergruppen erneut zu analysieren Dieses Vorgehen ist auch für die Beantwortung der Fragestellungen M.1. bis M.3. erforderlich.

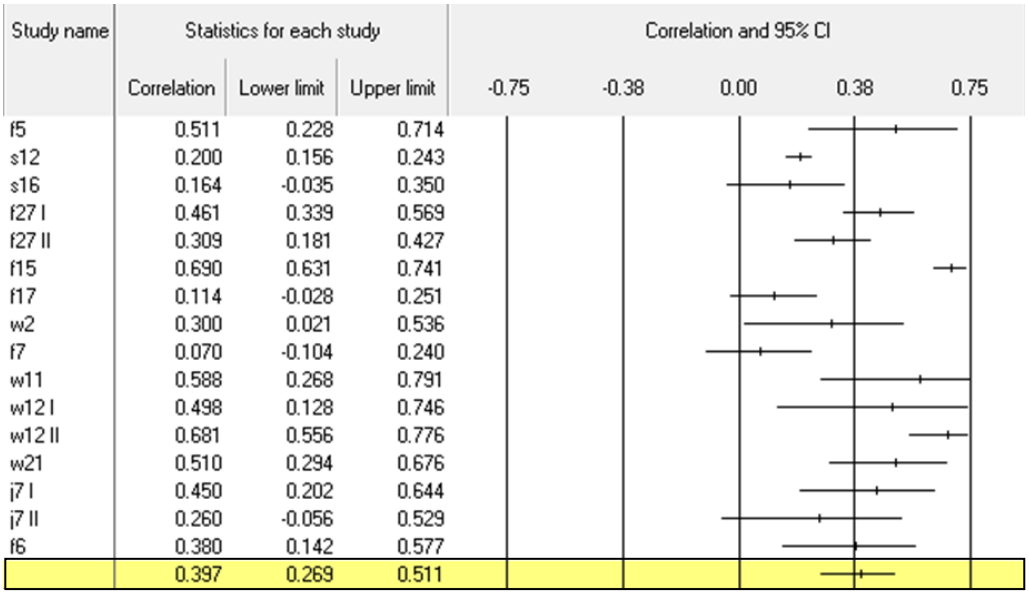

Abbildung 3: Metaanalyse: Übersicht der 16 Primärstudien

Bevor die einzelnen Fragestellungen konkret beantwortet werden, soll nachfolgend noch kurz auf die Überprüfung des *Publication Bias* eingegangen werden. Abbildung 4 zeigt den *Funnel-Plot* der 16 Studien. Man sieht, dass die Studieneffekte heterogen sind, sie liegen innerhalb und außerhalb des Dreiecks. Die gute methodische Qualität der Primärstudien wird hier ebenfalls sichtbar: Die Datenpunkte weisen insgesamt einen relativ geringen Standardfehler (*standard error*) auf. Der *Funnel-Plot* kann als eher symmetrisch angesehen werden und unterstützt damit nicht die These, dass ein starker *Publication Bias* vorliegt. Auf eine entsprechende statistische Testung wird an dieser Stelle verzichtet, da der Test auf den Prinzipien der Regression basiert und bei lediglich 16 Primärstudien die Power des Tests als verringert anzusehen ist (Sterne et al., 2011; Higgins & Green, 2011). Ein ähnliches Ergebnis lässt jedoch auch eine Berechnung des Kennwertes *Fail-safe-N* zu, der einen Wert von 1315 erreicht. Es müsste demnach 1315 nicht signifikante Studien zu dem

hier relevanten korrelativen Zusammenhang in die Analyse aufgenommen werden, damit der hier gefundene signifikante Effekt nicht mehr signifikant wird.

Im weiteren Verlauf der Analysen muss wegen der relativ geringen Studienanzahl der betrachteten Substichproben auf *Funnel-Plots* verzichtet werden, da sie nicht mehr sinnvoll einsetzbar sind (Sterne et al., 2011).

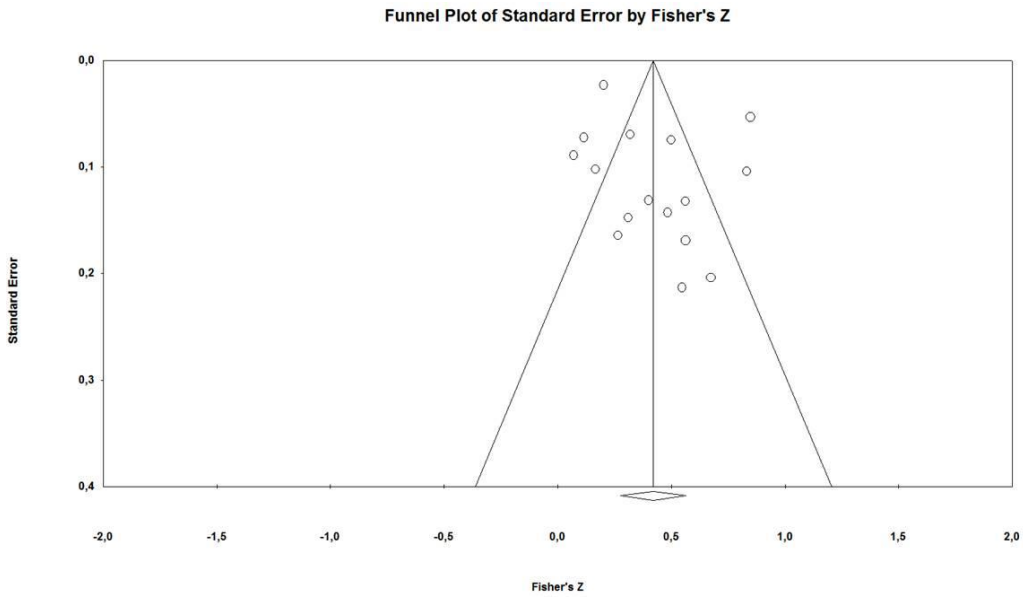

Abbildung 4: Metaanalyse: Funnel-Plot aller Primärstudien

Fragestellung M.1.

Bei Fragestellung M.1. geht es um den Zusammenhang zwischen DA_B und LK und dem Zusammenhang zwischen DA_A und LK. Hierzu wurden die Primärstudien danach unterschieden, ob ihr dynamisches Assessment auf arbeitsgedächtnis-, *reasoning*-, bzw. intelligenznahe Konstrukte (DA_A) oder auf die dynamische Lesekompetenz (DA_B) abzielt. Diese Aufteilung ist in Abbildung 5 zu sehen: Die ersten acht Studien erfassen DA_B, ihr Aggregat ist unmittelbar darunter abgetragen und umrahmt. Darunter befinden sich die Studien, die auf DA_A abzielen. Das entsprechende Aggregat dieser Studien kann der letzten Zeile der Abbildung entnommen werden. Der Unterschied zwischen beiden Gruppen beträgt .200 und wird nicht signifikant (p=.14).

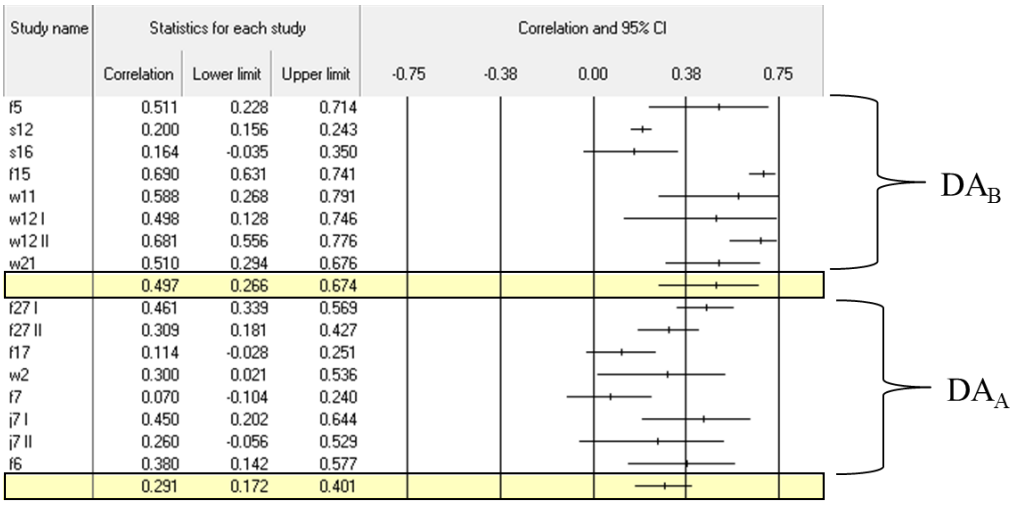

Abbildung 5: Metaanalyse: Korrelationen der LK mit DA_B und DA_A

Auf Grund der heterogenen Verteilung der Effekte in beiden Gruppen (DA_B: Q=159.171, df=7, p<.001, I^2=95.602; DA_A: Q=22.542, df=7, p=.002, I^2=68.947) werden die gefundenen Korrelationen von .497 und .291 jedoch nicht weiter verwendet, stattdessen sollen nun die Konfidenzintervalle der Korrelationen betrachtet werden. Eine grafische Inspektion dieser Konfidenzintervalle legt ebenso wie der nicht signifikante Test auf Unterschiede zwischen den Gruppen nahe, dass es sich hier nicht um zwei distinkte Gruppen handelt, die jeweils separat weiter analysiert werden müssen. Vielmehr sollen weiterführende Analysen diese beiden Gruppen aggregiert betrachten, auch um der inhaltlichen Breite der Konstrukte DA_A und DA_B Rechnung zu tragen.

Eine alternative Herangehensweise wäre eine weitere Untergliederung dieser beiden Gruppen anhand von möglichen Moderatorvariablen. Von einer solchen Fragmentierung soll nachfolgend jedoch abgesehen werden, da beide Gruppen nur jeweils acht Studien umfassen und jede weitere Unterteilung diesen ohnehin schon geringen Stichprobenumfang weiter reduzieren würde, was als nicht zielführend erachtet wird.

Fragestellung M.2.

Im Folgenden soll nun der Frage nachgegangen werden, welche LK-Facetten insbesondere mit DA_A und DA_B zusammen hängen (Fragestellung M.2.). Dazu

werden zunächst die Korrelationen aller 16 Primärstudien betrachtet und anschließend die Analysen jeweils für die DA_A-Studien und die DA_B-Studien wiederholt.

In diesem Zusammenhang ist insbesondere von Interesse, ob sich Unterschiede zwischen den Primärstudien finden lassen, die die hierarchiehohe, auf die Textebene abzielende Lesekompetenzfacette (LK_T) beziehungsweise die andere Lesekompetenzfacetten (LK_{NT}) als Untersuchungsgegenstand hatten.

Abbildung 6 stellt die entsprechenden Korrelationen aller 16 Primärstudien vor, jeweils gegliedert nach der statischen LK-Facette. Sie ist analog zu Abbildung 5 zu lesen. Zunächst sind fünf Studien abgetragen, die das Konstrukt „Texte verstehen“ (LK_T) erheben, alle anderen Studien sind darunter abgetragen. Die Aggregate der fünf Studien, die LK_T erheben und der elf Studien, die ein anderes Lesemaß erheben (LK_{NT}), sind jeweils umrahmt.

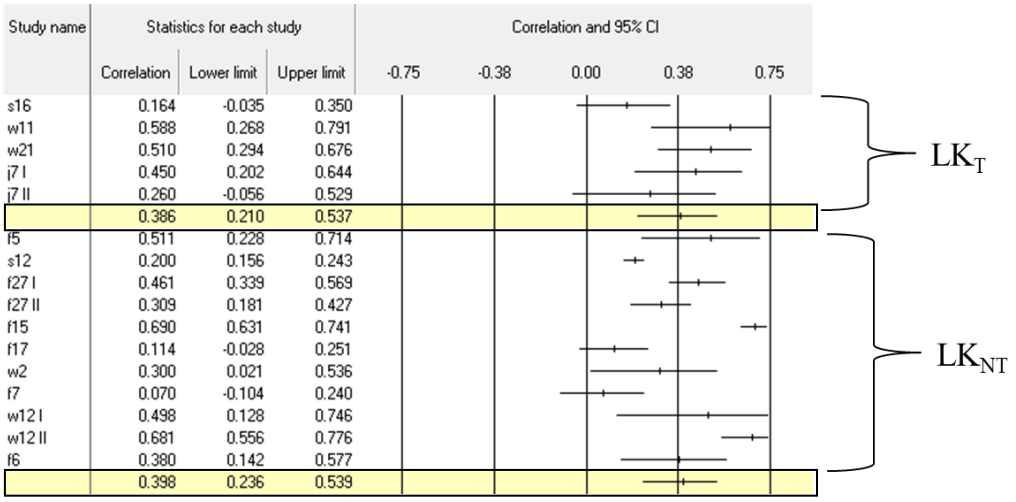

Abbildung 6: Metaanalyse: Korrelationen des DA mit LK_T und LK_{NT}

Die beobachteten Effektstärken der Studien, die LK_T erheben sind homogen (Q=9.403; df=4; p=.052; I^2=57.460). Damit ist eine sinnvolle Bewertung der gefundenen Ergebnisse zulässig. Als theoretische Legitimation dieser Analyse kann die starke Heterogenität der Lesekompetenz (vgl. Kapitel 3.1) betrachtet werden, die ihren Ausdruck in den unterschiedlichsten LK-Facetten (Kapitel 4.2.1) findet.

Die Korrelation zwischen LK_T und DA, das sich aus DA_B und DA_A zusammensetzt, ist positiv. Sie liegt bei .386 und ihr Konfidenzintervall hat die Grenzen .210 und .537. Die Studien, die ein anderes statisches Lesekompetenzmaß erhoben haben, sind bezüglich ihrer Effektgrößenverteilung nicht homogen (Q=172.985; df=10; p<.001; I^2=94.219).

Von den fünf Primärstudien, die LK_T erfassen, beziehen sich drei auf DA_B und zwei auf DA_A. Die drei DA_B-Studien berichten von einer durchschnittlichen Korrelation von .414 [.114; .644], während sich bei den beiden DA_A-Studien eine durchschnittliche Korrelation von .372 [.177; .539] finden lässt. Der Unterschied kann nicht als signifikant angenommen werden, da sich die entsprechenden 95%-Konfidenzintervalle überlappen.

Von den 11 Studien, die nicht explizit auf die Textebene abzielen (LK_{NT}), beziehen sich fünf auf DA_B und sechs auf DA_A. Beide Gruppen sind nicht signifikant voneinander verschieden (p=.13), mit einem durchschnittlichen Zusammenhang von .536 [.212; .754] bei den fünf DA_B-Studien und einem durchschnittlichen Zusammenhang von .273 [.133; .403] bei den sechs auf DA_A abzielenden Studien.

Damit ist die Forschungsfrage M.2. für die als LK_T bezeichnete hierarchiehohe Facette der Lesekompetenz, die auf die Ebene des Textes abzielt, hinreichend beantwortet. Für alle anderen Facetten (LK_{NT}) kann jedoch noch keine klare Antwort auf Forschungsfrage M.2. gegeben werden, weil die entsprechende Gruppe der Primärstudien keine ausreichende Homogenität aufweist. Das führt zu Forschungsfrage M.3.: Welche Variablen moderieren gegebenenfalls diesen Zusammenhang? Hierzu wurden systematisch verschiedene Variablen auf eventuell vorhandene Moderatoreffekte getestet.

Fragestellung M.3.

Ein erstes Ergebnis der systematischen Testung verschiedener Variablen auf Moderatoreffekte ist Abbildung 7 zu entnehmen. Hier sind nur noch auf LK_{NT} abzielende Studien berücksichtigt, also die 11 Primärstudien, die nicht die

Lesekompetenzfacette LK_T als statisches Lesemaß erhoben haben. Sie sind nach ihrer Stichprobenzusammensetzung unterteilt: die ersten beiden Zeilen bilden Studien ab, die eine maximal heterogene Zusammensetzung der Stichprobe umfassen. Dieser ist gemischt klinisch und nicht selektiert, d. h. sowohl Kinder mit einer klinischen Diagnose als auch Kinder aus dem normalen Leistungsspektrum finden sich in der untersuchten Population. Danach kommt die Gruppe mit einer rein klinischen Stichprobe (1 Studie) und ihr umrahmtes Aggregat und die Gruppe mit einer subklinischen Stichprobe (1 Studie) und ihr umrahmtes Aggregat. Subklinisch bedeutet in diesem Zusammenhang, dass zwar eine Minderleistung vorliegen kann beziehungsweise die untersuchten Kinder einer Risikopopulation angehören, jedoch keine feststehende Diagnose von einem Experten gestellt wurde. Die Gruppe mit einer gemischt subklinischen und nicht selektierten Stichprobe (7 Studien) und ihr Aggregat sind als letztes abgetragen.

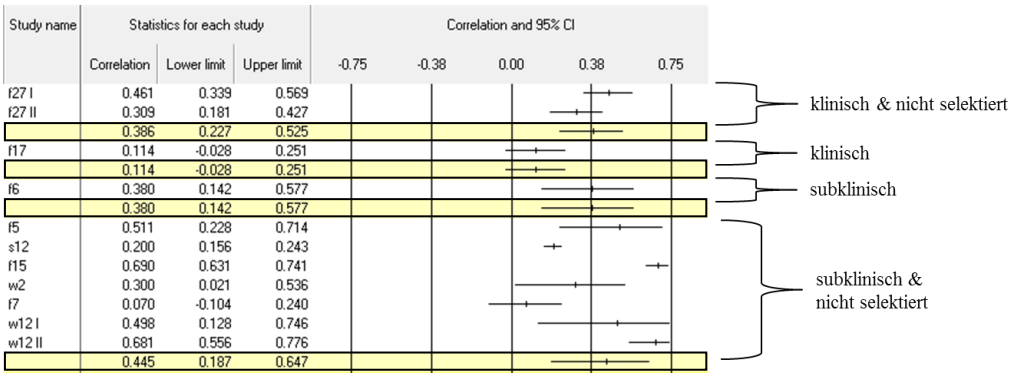

Abbildung 7: Metaanalyse: Korrelation der LK_{NT} mit dem DA in verschiedenen Populationen

Die beiden Studien mit gemischt klinischer und nicht selektierter Stichprobe können bezüglich ihrer Korrelationsverteilung als homogen angenommen werden (Q=3.098, df=1, p=.078, I^2=67.722). Sie zeigen eine durchschnittliche Korrelation von .386 zwischen dem DA und LK_{NT} und stammen aus derselben Publikation, ein Aspekt, der in der Interpretation der Metaanalyse (Kapitel 4.4) nochmals aufgegriffen und bewertet wird.

Die Gruppen, die nur aus einer Studie bestehen (f17 und f6), weisen trivialerweise als Aggregat den Wert dieser einzigen Studie auf. Der Test auf Homogenität der Effektstärkenverteilung kann damit nicht als sinnvolles Ergebnis im eigentlichen Sinne aufgefasst und weiter interpretiert werden. Die Varianz der Korrelationen der letzte Gruppe ist nicht homogen (Q=158.855, df=6, p<.001, I^2=96.223).

Die verbleibende Restkategorie, deren Effektstärkenverteilung nicht als homogen angesehen werden kann, umfasst damit noch neun LK_{NT}-Studien, die nicht über eine gemischt klinische und nicht-selektierte Stichprobe verfügen. Sie ist in Abbildung 8 dargestellt. Die Homogenität der Effektstärken ist nicht gegeben (Q=1166.720, df=8, p<.001, I^2=95.202). Auf eine weitere Fragmentierung dieser ohnehin schon recht kleinen Kategorie soll an dieser Stelle verzichtet werden, da sie in Hinblick auf die zu untersuchenden Fragestellungen nicht als sinnvoll erachtet wird. Es lässt sich jedoch festhalten, dass die wahre Korrelation als größer 0 angenommen werden kann, auch wenn sie nicht exakt bestimmbar ist.

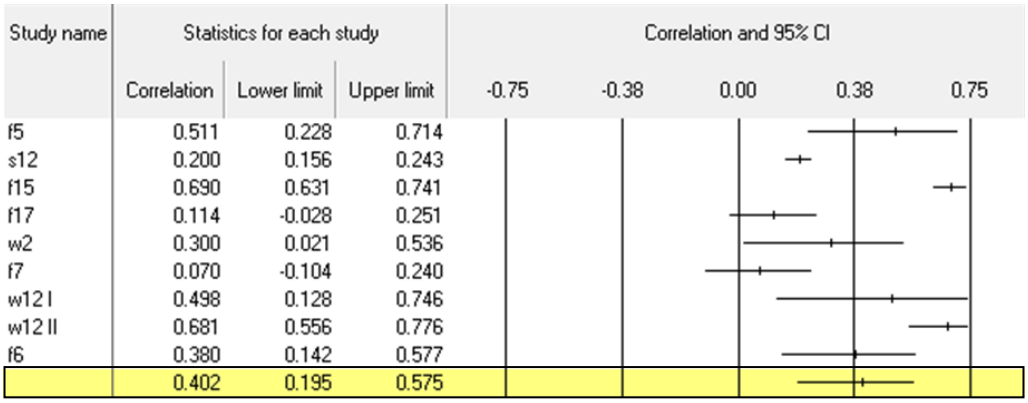

Abbildung 8: Metaanalyse: Korrelationen der LK_{NT} mit dem DA in Populationen ohne gemischt klinischer und nicht selektierter Stichprobenzusammensetzung

Zusammenfassung

Die Korrelation der statischen Lesekompetenz (LK) mit dem DA lässt sich nicht genau bestimmen. Sie kann jedoch als größer 0 angenommen werden und bewegt sich im Intervall zwischen .269 und .511.

Damit lassen sich zusammenfassend folgende Antworten auf die interessierenden Fragestellungen finden:

Fragestellung M.1.:

DA_B korreliert mit der statisch erfassten LK zwischen .266 und .674, der Zusammenhang ist positiv, ebenfalls positiv ist die Korrelation zwischen DA_A und der statisch erfassten LK, die zwischen .172 und .401 zu verorten ist (siehe Tabelle 9).

Tabelle 9: Zusammenfassende Beantwortung der Fragestellung M.1.

Korrelation [CI] mit der statischen Lesekompetenz	**Art des dynamischen Assessments**
positiv (.497 [.266; .674])	Bereichsspezifisches DA (DA_B)
positiv (.291 [.172; .401])	Allgemeines, nicht bereichsspezifisches DA (DA_A)

Fragestellungen M.2. und M.3.:

Für die Betrachtung der auf die LK-Facetten abzielenden Fragestellungen ist DA_A und DA_B stets zu DA zusammengefasst, um die jeweiligen Gruppen der Primärstudien nicht zu stark zu fragmentieren. Bei Studien, die auf LK_T abzielen, liegt die Korrelation mit dem DA bei .386. Bei Primärstudien, die LK_{NT} im Fokus ihrer Untersuchung hatten, findet sich ein bedeutender Moderatoreffekt: in Studien mit sehr heterogener Stichprobenbeschaffenheit zeigt sich eine durchschnittliche Korrelation von .386. Für alle anderen Primärstudien kann ein positiver Zusammenhang zwischen LK_{NT} einerseits und dem DA (DA_A und DA_B) andererseits angenommen werden (siehe Tabelle 10).

Tabelle 10: Zusammenfassende Beantwortung der Fragestellungen M.2. und M.3.

Korrelation [CI] der statischen Lesekompetenz mit dem dynamischen Assessment	**In Subpopulation mit Moderatorausprägung**
.386 [.210; .537]	Hierarchiehohes Lesekompetenzmaß (LK_T)
.386 [.227; .525]	Kein hierarchiehohes Lesekompetenzmaß (LK_{NT}) und eine gemischt klinische und nicht selektierte Stichprobe
positiv (.402 [.195; .575])	Restkategorie

4.4 Interpretation

Zusammenfassend kann damit festgehalten werden, dass ein positiver Zusammenhang zwischen LK und DA_B als gesichert angenommen werden kann. LK und im Besonderen LK_T teilen sich gemeinsame Varianz mit DA_A und DA_B.

Zur Überlegenheit des DA_B im Vergleich zum DA_A

DA_B korreliert in der Tendenz positiv und wohl auch stärker mit statischer Lesekompetenz als DA_A. Dies ist sicherlich durch die größere Überlappung der Konstrukte begründbar. Ein statischer Lesekompetenztest sollte inhaltlich näher an einem anderen auf das Lesen abzielenden Test sein als an einem anderen auf intelligenznahe Fähigkeiten abzielenden Test. Es liegt wahrscheinlich an der Anzahl der Studien, dass dieser Unterschied nicht signifikant wird. Ansatzweise deutet sich damit an, dass DA_B in Hinblick auf den Zusammenhang mit statisch erfasster LK dem nicht bereichsspezifischen DA_A überlegen ist. Die positive Korrelation kann jedoch nicht allein dem Ausmaß geschuldet sein, in dem Lesen im Rahmen des DA_B erfasst wird, sie muss vielmehr mit den Gemeinsamkeit von DA_A und DA_B zusammenhängen. Bei den DA_B-Studien und den DA_A-Studien ist aber insbesondere das dynamische Vorgehen eine hervorstechende Gemeinsamkeit. Der einzige Unterschied: DA_B ist bereichsspezifisch, erfasst also noch zusätzlich Lesen, dagegen erfasst DA_A nur kognitive Fähigkeiten. Damit hat die dynamische Komponente in der dynamischen Testung durchaus ihre Berechtigung als Prädiktor der statischen Lesekompetenz.

Zur Rolle des dynamischen Assessments beim statischen Lesekompetenzmaß LK_T

Dies wird bei der Betrachtung der Korrelationen des LK_T mit DA_A und DA_B ebenfalls deutlich. Die Korrelationen unterscheiden sich mit einer Differenz von .042 nicht signifikant voneinander. Die dem DA_A und DA_B gemeinsame dynamische Komponente spielt damit auch hier eine wichtige Rolle und kann die beobachteten Zusammenhänge besser erklären als die Komponente des Lesens im DA_B. Dies ist bei einem so hierarchiehohen Konstrukt wie LK_T, bei

dem kognitive Prozesse eine wichtige Rolle spielen, durchaus nachvollziehbar und wenig überraschend.

Ein alternativer Erklärungsansatz wäre, dass der Zusammenhang von LK_T mit DA_B ausschließlich auf die Komponente des Lesens und der Zusammenhang von LK_T mit DA_A ausschließlich auf das allgemeine kognitive Potential zurückzuführen sei. Dem muss entgegengehalten werden, dass auch die im dynamischen Test erfasste Lesekomponente nicht unabhängig vom allgemeinen kognitiven Potential sein kann (vgl. Kapitel 3.1.3.1). Kann also möglicherweise die Schnittmenge aus Lesen und allgemeinem kognitiven Potential die nahezu identischen Korrelationen verursachen? Diese Möglichkeit lässt sich nicht komplett ausschließen. Es ist jedoch eher weniger wahrscheinlich, dass diese Schnittmenge orthogonal zur Lernfähigkeits-komponente des dynamischen Tests ist.

Zur Rolle der systematischen Varianzeinschränkung in Hinblick auf hierarchiehohe (LK_T) vs. nicht hierarchiehohe Lesekompetenzfacetten (LK_{NT})

Ein weiterer Gesichtspunkt, der für die Relevanz der Lernfähigkeit spricht, ist der Moderatoreffekt der Stichprobenbeschaffenheit. Studien, die LK_{NT} erfassen sind in heterogener Population (gemischt klinische und nicht selektierte Stichprobe) in ihrer Korrelation mit DA vergleichbar mit Studien, die hierarchiehöheres LK_T erfassen. Dies ist bemerkenswert, da in einer solch heterogenen Population keine systematische Varianzeinschränkung vorliegt, die zu einer Unterschätzung der wahren Effektstärke führen würde, wie das möglicherweise bei den Ergebnissen in den leistungshomogeneren Populationen möglich ist. Das spricht generell für einen substantiellen, positiven Zusammenhang zwischen dem dynamischen Assessment und der statisch erfassten Lesekompetenz.

Dieser Befund impliziert außerdem, dass hier kognitive Faktoren eine Rolle spielen, die über den bereits bekannten Zusammenhang zwischen Intelligenz und Leseverständnis hinausgehen. LK_T ist im Vergleich zu LK_{NT} relativ „intelligenzlastig“ (vgl. Kapitel 4.2.1). Diese Eigenschaft bringt ihr aber keine

höhere Korrelation mit dynamischen Testverfahren ein, wenn man sie mit der Korrelation vergleicht, die LK_{NT} unter „optimalen Bedingungen“ erreichen kann. „Optimale Bedingungen“ meinen in diesem Fall eine Stichprobenbeschaffenheit, die nicht durch ihre systematische Einschränkung die tatsächliche Korrelation vermindert. Damit sind die gefundenen Zusammenhänge weniger auf die kognitiven Fähigkeiten zurückzuführen, die beim LK_T stärker zum Tragen kommen als beim LK_{NT}, also beispielsweise auf die Intelligenz oder auf das Arbeitsgedächtnis. Gleichzeitig müssen diesen Zusammenhängen jedoch kognitive Faktoren zu Grunde liegen, da sich sonst die Zusammenhänge mit LK_T zwischen dem DA_B und dem DA_A unterscheiden müssten. Es liegt nahe, dass die dynamische Komponente die hier ausschlaggebende kognitive Komponente ist.

Zu den beiden Studien mit gemischt klinischer und nicht-selektierter Stichprobe muss noch angemerkt werden, dass beide aus derselben Veröffentlichung (f27) stammen. Möglicherweise ist ihre Homogenität damit insbesondere dadurch zu erklären, dass beide Studien im selben Untersuchungsrahmen durchgeführt wurden.

Zwischenfazit

Die bisher diskutierten Befunde deuten insgesamt darauf hin, dass die dynamische Komponente der dynamischen Testung wohl als ein wichtiges Korrelat der statischen Lesekompetenz angesehen werden kann. Sie sind vergleichbar mit den metaanalytischen Befunden von Caffrey, Fuchs und Fuchs (2008), auch wenn deren empirische Herangehensweise sich teilweise stark von der hier vorgestellten Metaanalyse unterscheidet.

So wurden bei Caffrey et al. (2008) beispielsweise auch subjektive Lehrerurteile und hochgradig individualisierte Rückmeldungen in der Analyse berücksichtigt. Ein methodischer Hauptunterschied liegt daneben in der Art der Literaturrecherche begründet. Caffrey et al. (2008) durchsuchten drei Datenbanken nach möglichen Primärstudien. Dissertationsdatenbanken blieben hierbei unberücksichtigt. Die Rolle der Publikationsart der Primärstudien ist damit ein wichtiges Alleinstellungsmerkmal dieser Metaanalyse insbesondere

in Hinblick auf die Dichotomie bei der Literaturrecherche, bei der explizit Dissertationsdatenbanken berücksichtigt wurden. Nachfolgend soll dargestellt werden, dass dies die Validität der Metaanalyse jedoch nicht gefährdet.

Spezifika der Dissertationen

Insgesamt befinden sich unter den 16 Primärstudien fünf Doktorarbeiten. Auch wenn die Art der Publikation sicherlich wenig inhaltliche Relevanz aufweisen sollte, so sind doch einige methodische Gesichtspunkte finden, die Dissertationen von den anderen Primärstudien abgrenzen und sich damit auf die Validität der hier berichteten Ergebnisse auswirken können.

So stellt sich die Frage, ob die bei den Dissertationen gefundene Korrelation möglicherweise eine verminderte Validität aufweist und den Befunden aus in Fachzeitschriften publizierten Primärstudien mehr vertraut werden soll.

Wenn diese Überlegung ihre Richtigkeit hätte, dann müssten sich bei der Kodierung der Studien zwischen den Gruppen der Dissertationsschriften und anderen Publikationsformen Unterschiede in der methodischen Qualität finden. Dies war nicht der Fall. Die These einer verminderten Qualität auf Seiten der Dissertationen ist daher kritisch zu sehen. Vielmehr muss das Augenmerk auf die spezifischen Unterschiede zwischen diesen Publikationsarten gelegt werden. Strukturell unterscheiden sich Dissertationsschriften von *Papern* jedoch insbesondere durch eine wesentlich genauere Darlegung der methodischen Aspekte ihrer durchgeführten empirischen Studien. Damit ist tendenziell eine eher höhere Implementationskontrolle gegeben, die sich auch potentiell positiv auf die interne Validität der durchgeführten Untersuchung auswirken sollte. In *Papern* hingegen kann der Methodenteil relativ schnell und weniger tiefgehend abgehandelt werden.

Ein weiterer methodischer Unterschied zwischen *Papern* und Doktorarbeiten betrifft die Durchführungsobjektivität. Haben die Doktoranden selbst die Testung durchgeführt und sich anders verhalten als beispielsweise studentische Hilfskräfte, die als Testleiter eingesetzt wurden? Diese Unterschiede könnten systematische Auswirkungen auf die Motivation und auf das Verhalten der

jeweiligen Probanden haben, auch wenn diese sich nicht explizit in den Befunden niederschlagen. Sie legen jedoch keine Gefährdung der Validität nahe.

Daneben ist zu bedenken, dass sämtliche Dissertationen in speziellen Dissertationsdatenbanken gefunden wurden (vgl. Kapitel 4.2.2), die möglicherweise bestimmte Aufnahmekriterien für ihre Einträge haben, welche insbesondere bei internationalen Datenbanken nicht immer auf den ersten Blick ersichtlich sein müssen. Es kann nicht vollkommen ausgeschlossen werden, dass spezifische Datenbanken Anforderungen wie eine Mindesteffektstärke an ihre potentielle Einträge stellen oder andere datenbankspezifische Einflussgrößen dazu führen, dass die Dichotomie der Literaturrecherche Auswirkungen auf die dargestellten Befunde hat.

Des Weiteren ist noch anzumerken, dass es sich bei signifikanten Befunden stets auch um zufällige Effekte handeln kann und Signifikanztests lediglich eine obere Schranke für die Wahrscheinlichkeit darstellen, dass ein zufälliger Effekt fälschlicherweise als bedeutsam angenommen wird. Diese Problematik betrifft nicht nur die Dissertationsschriften, sondern jegliche hier berichteten Befunde.

Generell ist damit nicht von einer im Vergleich zur Analyse von Caffrey et al. (2008) verminderten Validität auszugehen.

Geringe Anzahl der Primärstudien

Ein in Hinblick auf die durchgeführten Signifikanztests für diese Metaanalyse bedeutsamer Kritikpunkt ist sicherlich die geringe Anzahl an Primärstudien. Diese wirkt sich auf die *Q*-Statistik und indirekt auch auf das Konfidenzintervall des I^2-Idexes aus (Huedo-Medina et al., 2006). Potentielle Moderatorvariablen werden damit eher nicht als Moderatorvariablen erkannt und auch Unterschiede zwischen Gruppen werden eher nicht signifikant. Sicherlich ist dieses Problem auch der Tatsache geschuldet, dass dynamisches Testen weniger stark beforscht wird als andere psychologische Themen (Sternberg & Grigorenko, 2002). Ein verstärktes Interesse der

wissenschaftlichen Psychologie an dynamischen Testverfahren wäre daher wünschenswert. Als eher unwahrscheinlich kann die Überlegung angesehen werden, dass bei der Literaturrecherche zu wenig potentielle Quellen für Primärstudien herangezogen wurden. Ein Vorteil der geringen Anzahl der Primärstudien ist jedoch, dass die in dieser Analyse verwendeten Studien als methodisch hochwertig und in sich relativ homogen angesehen werden können.

Generalisierbarkeit

Durch die starke Homogenität und die hohe methodische Qualität der Primärstudien kann die Generalisierbarkeit auf Studien mit ähnlichem Design und Fokus als gesichert angenommen werden.

Eine andere Frage betrifft die Generalisierbarkeit der Befunde auf andere Kulturen mit anderen Sprachen und Schriftsystemen. In dieser Metaanalyse wurden die Länder, in denen erhoben wurde, als potentieller Moderator mitberücksichtigt. Jedoch zeigte sich diese Moderatorvariable als nicht bedeutsam für die Gruppierung der Primärstudien. Da alle Studien in Ländern mit lateinischem Schriftsystem durchgeführt wurden, lassen sich die hier berichteten Ergebnisse nur auf solche Länder übertragen, die das lateinische Alphabet verwenden. Inwieweit dynamische Verfahren sensitiv für Lesemaße sind, denen eine andere Schrift zu Grunde liegt, bleibt dagegen eine offene Frage.

Damit können die Ergebnisse der metaanalytischen Untersuchung Gültigkeit für den zu konstruierenden dynamischen Lesekompetenztest haben, der ebenfalls auf dem lateinischen Alphabet basiert. Auch ist der Fokus des zu konstruierenden Lesekompetenztests hinreichend ähnlich zu den Primärstudien der Metaanalyse, in denen das dynamische Assessment dem interventionistischen Ansatz folgte und stets mittels standardisiertem Feedback realisiert wurde.

Ausblick

Eine weitere mögliche Fragestellung, die auf den hier vorgestellten Befunden der Metaanalyse aufbaut, zielt auf den Mehrwert der dynamischen

Komponente im Vergleich zu nicht dynamischen Verfahren ab: Ist der Zusammenhang der statischen Lesekompetenz (LK) mit dynamischen Tests vergleichbar mit dem Zusammenhang der statischen Lesekompetenz mit statischen Tests, die auf die allgemeinen kognitiven Fähigkeiten abzielen? Erste Hinweise lassen vermuten, dass das DA hier schwächere Korrelationen zeigt. So sind beispielsweise von Zusammenhängen der Lesekompetenz mit allgemeinen kognitiven Fähigkeiten auszugehen, die in der Größenordnung von bis zu r=.6 anzusiedeln sind (vgl. Kapitel 3.1.3.1). Ein systematischer metaanalytischer Vergleich wurde bislang jedoch noch nicht realisiert. Die Varianzaufklärung, die die dynamische Komponente des dynamischen Assessments in der statischen LK bringt, hängt natürlich auch von der Stärke ab, mit der die Lesekomponente im DA_B mit der statischen Lesekompetenz (LK) zusammenhängt.

Damit leiten sich weitere Fragestellungen ab: Wie ist der Zusammenhang zwischen DA_B und der im Rahmen des dynamischen Assessments erhobenen dynamischen Komponente? Anders ausgedrückt: Wieviel „Dynamik“ und wieviel „Lesen“ steckt im DA_B? Erste Recherchen deuten darauf hin, dass es hierzu bislang noch nicht ausreichend viele Untersuchungen gibt. Dennoch würde die Klärung dieser Frage helfen, den exakten Mehrwert an Varianzaufklärung zu bestimmen, den die Lernfähigkeitskomponente des dynamischen Assessments liefert.

So können die vorliegende Studie und ihre geplanten Folgeuntersuchungen auch als Antwort auf den Vorwurf zu sehen sein, dass das dynamische Assessment bislang zu wenig durch Verbindungen zu anderen Forschungsgegenständen legitimiert ist (vgl. Kapitel 2.6). Hier wird sicher nicht der Anspruch erhoben, diesen Kritikpunkt vollständig zu entkräften. Jedoch kann die hier vorgestellte metaanalytische Zusammenschau einen Teil dazu beitragen, das Konzept des dynamischen Assessments besser zu durchdringen.

Davon ausgehend würden sich auch spezifische Implikationen für die Förderung der Lesekompetenz ableiten lassen. So deutet sich bereits jetzt an,

dass das dynamische Assessment einen wertvollen und eigenständigen Beitrag zur Vorhersage und wohl auch zur Verbesserung von Lesekompetenz leisten kann. Weitere Studien zur statischen Lesekompetenz sollten daher auch ein besonderes Augenmerk auf die Lernfähigkeit legen.

Die Entwicklung eines dynamischen Lesekompetenztests kann damit als ein Projekt angesehen werden, welches neben theoretischen Überlegungen auch durch ein solides empirisches Fundament als vielversprechend angesehen werden kann. Die konkrete Umsetzung der Testentwicklung soll die zentralen Ergebnisse der metaanalytischen Untersuchung angemessen berücksichtigen.

4.5 Implikationen für die angestrebte Testentwicklung

Aus den Befunden der Metaanalyse folgt damit für die Testentwicklung, dass die dynamische Komponente und die Lesekompetenz nicht als hinreichend identisch angesehen werden können, um zu einem gemeinsamen Index aggregiert werden zu können. Vielmehr müssen zwei Indizes entwickelt werden, einer, der die statisch erfasste Lesekompetenz abbildet und einer, der die dynamische Komponente abbildet.

Die Befunde der metaanalytischen Untersuchung legen darüber hinaus nahe, dass die dynamische Komponente und die Lesekompetenz keine zueinander orthogonalen Konstrukte sind. Dies wird wohl auch dem Einfluss allgemeiner kognitiver Fähigkeiten geschuldet sein. Des Weiteren werden im zu konstruierenden dynamischen Lesekompetenztest beide Indizes aus demselben Testmaterial abzuleiten sein, womit eine weitere Konfundierung der beiden Konstrukte miteinander nicht ausgeschlossen werden kann.

Bei der Entwicklung der beiden Indizes für die Lesekompetenzkomponente und die dynamische Komponente muss daher eine möglichst minimale Überlappung angestrebt werden, um den inkrementellen Mehrwert der dynamischen Testversion gegenüber einer Testversion ohne dynamische Komponente zu maximieren. Die Herleitung eines Kennwerts der dynamischen Komponente wird in Kapitel 8.3.2.3 unternommen. Sie beruht auf den Items des Tests, die primär die Lesekompetenz erfassen sollen. Die Herleitung und Entwicklung dieser Testmaterialien soll im nächsten Kapitel dargelegt werden.

5 Entwicklung der Materialien des dynamischen Lesekompetenztests

5.1 Ausgangssituation und Vorüberlegungen

Bei der Entwicklung der Testmaterialien werden die bisherigen Befunde und Überlegungen zum dynamischen Assessment und zur Lesekompetenz miteinander verbunden. Wie in Kapitel 2.3 und Kapitel 2.5 dargelegt bietet sich zur Umsetzung des dynamischen Lesekompetenztest eine computeradministrierte Testung im Multiple-Choice-Format an, die auch gruppenweise in den jeweiligen Schulklassen erfolgen kann. Aufgaben im Multiple-Choice-Format können als qualitativ hochwertig und für Computertestungen geeignet angesehen werden. Sie sind in diagnostischer Hinsicht vergleichbar mit Aufgaben mit offenem Antwortformat (Lindner, Strobel & Köller, 2015). Das Format kommt beispielsweise bei der Erfassung der Leseleistung im Rahmen internationaler Schulleistungsstudien zum Einsatz (z. B. PISA, Prenzel, Carstensen, Frey, Drechsel & Rönnebeck, 2007, S. 48; IGLU, Bos et al., 2004, S. 52). Dabei bietet sich das besonders ökonomische *train-within-test*-Design an, bei dem standardisiertes Feedback gegeben wird. Für die Erstellung der Testaufgaben kommt den Ergebnissen von Kapitel 3.1.3.2 besondere Bedeutung zu, in dem aufgezeigt wurde, welche Textmerkmale sich in welcher Weise auf die Lesekompetenz der Kinder auswirken.

Technische Rahmenbedingungen

Der zu konstruierende dynamische Lesekompetenztest knüpft an die Arbeit von Golke et al. (2015) an, in der im Bereich der Sekundarstufe I ein computeradministrierter Lesekompetenztest im Multiple-Choice-Format mit dynamischer Komponente erfolgreich implementiert werden konnte. Den Schülern wurden am Computer innerhalb einer speziell dafür entwickelten grafischen Benutzeroberfläche Texte präsentiert, zu denen jeweils mehrere Fragen gestellt wurden. Die richtige Antwort konnte ausgewählt werden aus einer Auswahl von Antwortalternativen. Bei einer falschen Antwort erfolgte

Feedback, das von einem *pedagogical agent* gegeben wurde. Der *pedagogical agent* ist eine computeranimierte Figur, die die Schüler durch das Programm begleitet. Bei richtigen Antworten meldete die Figur entsprechend den Erfolg zurück. Der Einsatz eines *pedagogical agents* ist insbesondere förderlich für die Motivation des Lerners (Johnson & Rickel, 2000). Verschiedene Ergebnisse einer Übersichtsarbeit legen nahe, dass der *pedagogical agent* für das hier beschriebene Projekt geeignet ist. Erstens eignen sich *pedagogical agents* insbesondere für jüngere Lerner ab der Primarstufe, weniger jedoch für ältere Lerner. Die positiven Auswirkungen auf die Motivation scheinen speziell bei jüngeren Lernern besonders ausgeprägt zu sein. Zweitens scheinen speziell dann positive Auswirkungen auf den Lerner vorzuliegen, wenn der *pedagogical agent* eine humanoide Form aufweist. Drittens hat sich der Einsatz der *pedagogical agents* insbesondere in Felderhebungen bewährt, bei denen direkt in der Schule getestet wurde (Schroeder, Adesope & Gilbert, 2013). Damit kann der Einsatz eines humanoiden *pedagogical agents* bei in der Schule stattfindenden Erhebungen an Kindern dritten und vierten Klassenstufen als generell sinnvoll angesehen werden. Die Autoren der Übersichtsarbeit gehen außerdem davon aus, dass die erlebte Interaktion mit dem *pedagogical agent* insbesondere stark auf jüngere Lerner wirkt. Dieser Effekt kann teilweise auch auf das vom Lerner oft aufgebaute persönliche Interesse am *pedagogical agent* zurückgeführt werden (Veletsianos & Russell, 2013), welches als reziprok erlebt wird. So können *pedagogical agents* den Eindruck erwecken, sich für den Lerner und seinen Fortschritt zu interessieren, was positive Auswirkungen auf dessen Motivation hat (Johnson & Rickel, 2000). Das zeigt auf, dass der Einsatz eines *pedagogical agents* sich nicht nur auf die Ebene der reinen Informationsübermittlung reduzieren lässt, sondern auch motivationale Unterstützung bietet. Des Weiteren wurde im hier beschriebenen Projekt ein Belohnungssystem als zusätzliche Motivationsquelle (Rudolph, 2003, S. 46) eingesetzt. Für jede richtige Antwort sammelten die Schüler Punkte in Form von Goldmünzen, die sie jederzeit auf dem Bildschirm sehen konnten.

Der zu entwickelnde dynamische Lesekompetenztest soll in diese bereits vorhandene Softwareumgebung eingebettet werden. Nachfolgend soll der

schematische Aufbau des Programms beschrieben werden. Im oberen Teil des Bildschirms befindet sich ein Textfeld, in welchem der Aufgabenstamm steht. Der untere Teil der Anzeige ist zweigeteilt. Links stehen die Frage und die Antwortalternativen. Auf der rechten Seite befindet sich der *pedagogical agent* und daneben die bislang gesammelten Goldmünzen und der Punktestand. Oberhalb befindet sich nochmal ein Textfeld. In diesem wird die Rückmeldung auf falsche Antworten gegeben.

Das Computerprogramm wurde in JAVA geschrieben, welches durch seine Plattformunabhängigkeit große Vorteile im Einsatzfeld Schule bietet. Unabhängig von den an den Schulen vorhandenen Betriebssystemen kann das JAVA-Tool verwendet werden, auch ohne vorab auf den Rechnern der Schule ein Programm installieren oder Zugang zum Internet sicherstellen zu müssen. Das Programm wird auf Wechseldatenträgern gespeichert und direkt auf diesen aufgerufen. Die während der Laufzeit eingegebenen Daten, die vom Probanden ausgewählten Antworten und die jeweiligen Reaktionszeiten werden für jeden Programmdurchlauf in einer xml-Datei gespeichert.

Weiterentwicklung und Adaption formaler Testaspekte

Ähnlich wie in der Arbeit von Golke et al. (2015), bei der Inferenzen und Situationsmodelle im Fokus standen, soll auch in dem in dieser Arbeit zu konstruierenden Test das Augenmerk auf implizite Aufgaben gelegt werden. Für eine Anwendung im Primarbereich sind jedoch einige Anpassungen des Testmaterials nötig, die insbesondere auf ein im Vergleich zu dem von Golke verwendeten Test vermindertes Schwierigkeitsniveau abzielen. Diese betreffen unter anderem den Aufgabenstamm der Items und sollen nachfolgend kurz skizziert werden. Die theoretische Basis dieser Modifikationen ist in Kapitel 3.1.3.2 dargestellt, in welchem neben den Auswirkungen der Textsorte und der Textlänge und auch die Auswirkungen der Anzahl unbekannter Wörter im Text auf die Leseleistung aufgezeigt wurden.

Die Anzahl der Distraktoren war bei Golke et al. (2015) konstant vier, für den Einsatz im Primarbereich soll sie auf drei reduziert werden. Drei Distraktoren können als ausreichend angesehen werden, gleichzeitig wird durch die

Reduktion der Distraktorenanzahl das Schwierigkeitsniveau vermindert (Eid & Schmidt, 2014, S. 107). Sie sind darüber hinaus so zu konstruieren, dass sie ein gewisses Maß an Plausibilität aufweisen, um überhaupt als mögliche richtige Antwort in Erwägung gezogen zu werden. Je attraktiver hierbei die Distraktoren sind, umso schwieriger sollte die Testaufgabe für den Probanden sein. Die Distraktoren sollen also nach Möglichkeit im Aufgabenstamm vorkommen und ebenso wie die richtige Antwort mit der Fragestellung assoziiert werden können, beispielsweise durch semantische Ähnlichkeit oder einer ähnlichen Funktion im Situationsmodell. Je stärker dabei die Assoziation mit der Fragestellung, umso attraktiver sollte der Distraktor sein und umso weniger wird eine richtige Antwort des Probanden zu erwarten sein. Daher kommt nicht nur der Anzahl sondern auch der Konstruktion der Distraktoren in Hinblick auf die Aufgabenschwierigkeit Bedeutung zu.

Daneben hat Golke im Aufgabenstamm mit relativ langen Texten gearbeitet, die allein durch ihre Länge höhere Anforderungen an den Leser stellen. Wie in Kapitel 3.1.3.2 dargelegt, ziehen längere Texte verglichen mit kürzeren Texten eine verminderte Leseleistung nach sich. In Hinblick auf die Zielpopulation, zu der auch Kinder mit spezifischem Förderbedarf gehören, ist es folglich angebracht, die Textlänge derart zu reduzieren, dass Motivation und Aufmerksamkeit der Kinder nicht unnötig beeinträchtigt werden. Daher soll jeder Aufgabenstamm höchstens 40 Wörter enthalten. Eine weitere aus Kapitel 3.1.3.2 stammende Implikation für die Testkonstruktion betrifft Worte, die der Zielpopulation tendenziell eher unbekannt sein dürften. Solche Worte sollen möglichst vermieden werden. Generell ist eine kindgerechte Formulierung der Items anzustreben. Darüber hinaus sollen die Items so formuliert werden, dass sie über eine möglichst lang anhaltende Aktualität verfügen und keine Wertungen expliziter oder impliziter Natur beinhalten (Pospeschill, 2010).

Um eine Konfundierung zwischen den Items auszuschließen, soll darüber hinaus bei der Konstruktion des Aufgabenstamms außerdem darauf geachtet werden, dass keine zwei Items denselben Aufgabenstamm haben, sich also auf denselben Text beziehen. Die Themen der Texte sollen überdies eine möglichst große Vielfalt aufweisen. Hierbei ist auch darauf zu achten, dass die

Themenauswahl den kindlichen Interessen nahe kommt. Ein empirischer Überblick über die Interessen von Kindern im Primarbereich findet sich beispielsweise bei Pruisken (2005) und auch bei Groenwald (2012) werden Hinweise auf mögliche Themen gegeben, die für Kinder im Grundschulalter interessant sein können.

Als weiteres Merkmal des Aufgabenstamms, an Hand dessen die zu entwickelnden Items systematisch variiert werden sollen, ist die Textart zu nennen, die, wie in Kapitel 3.1.3.2 dargelegt, mit dem Vorwissen des Lesers einerseits und der Notwendigkeit, Inferenzen zu ziehen, andererseits in Zusammenhang steht. Analog zu etablierten Lesetests wie dem Frankurter Leseverständnistest für 5. und 6. Klassen (Souvignier, Trenk-Hinterberger, Adam-Schwebe & Gold, 2008) oder dem Hamburger Lesetest für 3. und 4. Klassen (Lehmann et al., 2006) sollen sowohl narrative Texte, als auch Sachtexte zum Einsatz kommen. Dies ist in Einklang mit den Schlussfolgerungen aus Kapitel 3.1.3.2, wonach davon ausgegangen werden kann, dass die Textart per se keine direkte Auswirkung auf den Schwierigkeitsgrad des Textes hat. Vielmehr sind die beim Lesen ablaufenden Prozesse miteinander vergleichbar. Dabei sind in Hinblick auf die Rolle des Vorwissens bei narrativen Texten und Sachtexten unterschiedliche Schwerpunkte zu wählen. Die narrativen Texte sind so zu gestalten, dass sie kein spezielles Vorwissen in bestimmten Bereichen benötigen, sondern Situationen darstellen, die aus der Alltagswelt der Kinder kommen und ihnen aus dem täglichen Leben bekannt sind. Bei der Inferenzbildung soll der Schwerpunkt auf Sachtexten liegen, die weniger vielschichtig sind und eindeutigere Inferenzen zulassen. Der thematische Schwerpunkt der Sachtexte ist im naturwissenschaftlichen Bereich anzusiedeln. So kann der Einfluss kulturbedingten Vorwissens zwar nicht komplett ausgeschaltet, jedoch minimal gehalten werden.

5.2 Theoretische Einbettung des zu konstruierenden Materials

Die bislang aufgeführten Vorüberlegungen spannen den Rahmen auf, in dem die systematische Itemgenese erfolgen kann. Bevor die Items generiert werden können, muss jedoch noch genauer präzisiert werden, welche inhaltlichen Aspekte der Lesekompetenz die Items abbilden sollen (Art der erfragten Information) und wie sich die Fragen nach diesen Informationen gegebenenfalls systematisch voneinander unterscheiden (Art der Aufgabe). Als Grundlage dieser für die Inhaltsvalidität essenziellen Spezifizierung dienen die theoretischen Aspekte und empirischen Befunde der Leseforschung (vgl. Kapitel 3).

Art der erfragten Information

Wie in Kapitel 3 dargelegt, ist es für den Leseprozess von großer Bedeutung, Inferenzen ziehen und Situationsmodelle generieren und anpassen zu können. Die einschlägige Literatur berichtet von vielen verschiedenen Inferenzen (vgl. Kapitel 3.2), von denen lediglich eine Auswahl in diesen Test aufgenommen werden kann.

Spezifisch soll der zu konstruierende Lesetest auf Inferenzen zu Raum, Zeit und Kausalität abzielen, die zum einen als empirisch gesichert angesehen werden können (z. B. Zwaan, Langston & Graesser, 1995; Graesser, Singer & Trabasso, 1994) und zum anderen durch das *event-indexing model* eine valide theoretische Basis haben (vgl. Kapitel 3.4). Somit können auch Dimensionen erster Ordnung (Raum und Zeit) sowie Dimensionen zweiter Ordnung (Kausalität) als im zu entwickelnden Lesekompetenztest hinreichend berücksichtigt und umgesetzt angesehen werden. Die Dimensionen des Protagonisten und der Intention finden keinen Eingang in den Lesetest, da sie bei Sachtexten nicht sinnvoll umgesetzt werden können. Raum, Zeit und Kausalität sind außerdem durch ihre Eigenschaft als relativ grundlegende Konzepte als für den Primarbereich generell geeignet anzusehen. Gleichzeitig verfügen sie jedoch auch über eine hinreichend hohe Komplexität, so dass nur automatisch ablaufende Inferenzbildungen als weniger wahrscheinlich angesehen werden können. Um Wiederholungseffekten vorzubeugen und eine

hinreichende Itemvariabilität, beispielsweise in Hinblick auf die Schwierigkeit der Aufgabe, zu erhalten, wird die Art der abzuprüfenden Information innerhalb jeder Inferenzart variiert. Beispielsweise können lokale Inferenzen nicht nur auf absolute Positionen der Entitäten in einem bestimmten Raum abzielen, sondern auch auf relative Positionen distinkter Entitäten zueinander.

Dem Modell von Kintsch zu Folge ist das Ziehen von Inferenzen eine elementare Voraussetzung für die Bildung einer adäquaten Repräsentation (vgl. Kapitel 3.3). Um zu überprüfen, ob diese grundlegenden Voraussetzung vorhanden ist, soll noch eine zusätzliche Inferenzart Berücksichtigung finden: Brückeninferenzen, die beispielsweise für die Verknüpfung von Propositionen nach dem *construction-integration-model* fundamental sind (Kapitel 3.3). Diese können als besonders leicht zu generieren angesehen werden. Sie sollen stets gleich abgeprüft werden und immer auf das korrekte Zuordnen eines Personalpronomens zu einer Entität abzielen. Damit soll zum einen gewährleistet sein, dass sich die Aufgaben zu den Brückeninferenzen in ihrer Schwierigkeit nicht fundamental unterscheiden. Gleichzeitig wird durch diese Standardisierung eine Art „*Baseline*“ erzeugt, mit deren Hilfe diejenigen Kinder im Primarbereich erkannt werden können, deren Leseschwierigkeiten solcherart sind, dass sie nicht über die für die Inferenzbildung minimal nötigen Kompetenzen verfügen. Damit werden neben den relativ elaborierten Inferenzen der maximalistischen Inferenztheorie auch jene Inferenzen berücksichtigt, welche nach der in Kapitel 3.2 beschriebenen minimalistischen Hypothese automatisch gebildet werden (Harley, 2008, S. 370). Des Weiteren sollten diese besonders einfachen Aufgaben durch die leicht zugänglichen Erfolgserlebnisse besonders positiv auf die Motivation der Schüler wirken, die im Bereich der Lesekompetenz geringere Leistungsfähigkeit vorweisen können und dort seltener Erfolg erleben. Diese Überlegungen sind insbesondere in Hinblick auf Kinder mit spezifischem Förderbedarf interessant, die eine besondere Zielgruppe des zu konstruierenden Verfahrens darstellen und auf die in Kapitel 8.2 und in Kapitel 8.4 nochmal gesondert eingegangen wird.

Art der Aufgabe

In Hinblick auf die Ausführungen zum Kompetenzstufenmodell (Kapitel 3.5) scheinen sich für die bei kurzen Texten an den Leser gestellten Anforderungen insbesondere drei Arten von Aufgaben anzubieten: Aufgaben mit expliziter Textinformation, Aufgaben mit paraphrasierter Textinformation und Aufgaben mit impliziter Textinformation (vgl. Tabelle 2).

Aufgaben mit expliziter Textinformation stellen die geringsten Anforderungen an den Leser. Die für die Lösung des Items benötigte Information steht wortwörtlich im Aufgabentext und muss nur noch gefunden werden (Anforderung: *locate*). Dafür wird kein spezifisches Vorwissen zum Textinhalt benötigt. Im Folgenden wird diese Anforderung Lokalisieren genannt.

Anspruchsvoller ist der Prozess, der zur richtigen Lösung der Aufgaben mit paraphrasierter Textinformation führt. Der Leser muss die relevante Textinformation nicht nur lokalisieren können, sondern auch erkennen, dass sie synonym zur richtigen Lösung ist und diese Synonyme ineinander überführen. Dieser nachfolgend als Paraphrasieren (*paraphrase*) bezeichnete Prozess benötigt Vorwissen über die synonyme Umschreibung der konkreten sprachlichen Formulierung im Aufgabenstamm.

Aufgaben mit impliziter Textinformation (Anforderung: *integrate*) stellen die höchsten Anforderungen an den Leser. Eigenes Vorwissen muss in Bezug zu den vorhandenen Textinformationen gesetzt und zu einer kohärenten mentalen Repräsentation integriert werden. Diese Leistung soll im Folgenden als Integrieren bezeichnet werden.

Mit diesen drei Aufgabenarten wird der Versuch unternommen, das zum Lösen der Aufgabe benötigte Vorwissen systematisch zu variieren. Damit kann der Einfluss des Vorwissens auf die Testleistung (vgl. Kapitel 3.1.3.1) zwar nicht vollständig kontrolliert werden, er ist nun jedoch zumindest in Teilen determiniert. Während das Vorwissen des Lesers bei impliziten Aufgaben benötigt wird, spielt es bei expliziten Aufgaben weniger eine Rolle.

Kombination von Informations- und Aufgabenart

Für eine Verknüpfung der zu erfassenden Informationsart mit der Art der Aufgabe soll jede Informationsart mit jeder Aufgabenart kombiniert werden. Dabei ist eine Besonderheit zu beachten: Implizite Informationen können nicht valide durch die im Lesetest zum Einsatz kommenden Brückeninferenzen abgefragt werden. Grund hierfür ist, dass die zu erschließende Information in der Dimension Kausalität anzusiedeln ist. Daher können nur explizite und paraphrasierte Items zum Einsatz kommen, wenn die zu erfragende Information aus dem Bereich der Brückeninferenzen stammt (in dieser Arbeit „Brückeninformation“ genannt). Nur bei der Abfrage lokaler, temporaler und kausaler Textinformation werden daher implizite Aufgaben zum Einsatz kommen. Es gibt damit insgesamt 11 mögliche Kombinationen von Aufgaben- und Informationsart.

Tabelle 11: Übersicht über die Aufgabenarten der Testitems

Textinformation im Aufgabenstamm	**Merkmal der Aufgabe**	**Anforderung an den Leser**	**Vorwissen benötigt?**	**Anwendung bei folgenden Informationsarten**
explizit	Lösung steht wortwörtlich im Text	lokalisieren	nein	lokale Information, temporale Information, kausale Information, Brücken-information
paraphrasiert	Lösung der Aufgabe (X) steht mit anderen Worten (Y) im Aufgabenstamm	paraphrasieren	ja, benötigt Vorwissen, dass X und Y bedeutungsgleich sind	lokale Information, temporale Information, kausale Information, Brücken-information
implizit	Lösung steht nicht im Aufgabenstamm, sondern muss erschlossen werden	integrieren	ja, benötigt Vorwissen zum Situationsmodell des Aufgabenstamms	lokale Inferenz, temporale Inferenz, kausale Inferenz

Einen Überblick über die drei Aufgabenarten, die im zu konstruierenden Lesekompetenztest entwickelt werden, ihren Bezügen zu bereichsspezifischen Vorwissen und ihren Anwendungen bei den verschiedenen Informationsarten findet sich in Tabelle 11. Ihre Schwierigkeit lässt sich nach einem von Kirsch (2001) entwickelten Schema abschätzen, wonach neben der Aufgabenart auch die Informationsart Auswirkungen auf den Aufwand haben soll, den der Leser bei der Bearbeitung der Aufgabe hat: Lokale Items gelten als tendenziell einfacher als temporale Items. Diese sind wiederrum tendenziell einfacher als kausale Items (Kirsch, 2001). Lokale Items können darüber hinaus als in der Tendenz anspruchsvoller als die Items angesehen werden, die auf die fundamentaleren Brückeninformationen abzielen.

Die zu erstellenden Aufgaben sollen sich somit insgesamt in drei Dimensionen systematisch unterscheiden: der Art der Textinformationen, auf die diese Aufgabe abzielt (lokale, temporale oder kausale Informationen bzw. Brückeninformationen), die konkrete Anforderung, die die Aufgabe an den Testanden stellt (explizite, paraphrasierte oder implizite Aufgaben) und der Textart des Aufgabenstamms (narrativer Text oder Sachtext). Damit ergibt sich ein Itemschema, welches genau drei systematische Variationen beinhaltet. Ausgehend hiervon kann damit eine einheitliche Nomenklatur für die Items entwickelt werden (Kapitel 5.3), die auch im Folgenden beibehalten werden soll.

Art der dynamischen Komponente

Die Umsetzung der dynamischen Komponente des dynamischen Lesekompetenztests soll an die zentralen Punkte des Kapitels 2 anknüpfen. Sie soll aus psychometrischen Gründen interventionistisch orientiert sein (vgl. Kapitel 2.4.1) und auf Veränderungen abzielen, die durch Intervention evoziert werden. In Anlehnung an die Arbeit von Golke et al. (2015) soll die Testung im *train-within-test*-Format gehalten sein und damit die *baseline reserve capactiy* erfassen (vgl. Kapitel 2.4.2). Im Unterschied zum Ansatz von Guthke (Kapitel 2.5) wird bei Golke et al. (2015) nur eine Hilfestellung nach einer falschen Antwort und damit nur zwei Versuche pro Aufgabe gegeben. Dies soll für den zu konstruierenden Test beibehalten werden. Dadurch soll auch dem

Kritikpunkt der verminderten Ökonomie der dynamischen Testung teilweise begegnet werden (Kapitel 2.6), es muss unter anderem pro Aufgabe nur eine spezifische Rückmeldung entwickelt werden. Die Responsivität auf die gegebenen Hilfestellungen, welche sich im Antwortverhalten im zweiten Versuch niederschlägt, soll somit dem Grundgedanken des dynamischen Assessments Rechnung tragen. Ihre genaue Implementierung ist für die konkreten Schritte der Materialentwicklung jedoch von untergeordneter Bedeutung und wird in Kapitel 8.3.2.3 thematisiert.

Für die dynamische Komponente ist neben den Ergebnissen aus Kapitel 2 auch von Relevanz, welche Art der Rückmeldung sich bei Lesekompetenzaufgaben empirisch besonders bewährt haben. Im Rahmen der Arbeit von Golke hat sich hierbei insbesondere elaboriertes Feedback als besonders hilfreiche Art der Rückmeldung erwiesen, bei dem neben einer Performanzrückmeldung auch eine Hilfestellung gegeben wird, wie die in der Aufgabe geforderte Leistung (z. B. das Suchen einer Information im Text, das Ziehen einer bestimmten Inferenz) erbracht werden kann (Golke et al., 2015; Golke, 2013). Daher soll diese Art des Feedbacks auch in dem hier zu erstellenden Lesekompetenztest Anwendung finden.

5.3 Nomenklatur der Testitems

Jedes Item erhält eine Bezeichnung nach dem Schema *XYZ0*. Dabei gilt:

X: bezeichnet die Aufgabenart: E/P/I steht für explizite/paraphrasierte/implizite Information

Y: bezeichnet die Informationsart: B steht für Brückeninformationen, L/T/K für lokale/temporale/kausale Informationen

Z: bezeichnet die Textart: S steht für Sachtexte, N steht für narrative Texte

0: Zahl: Durchnummerierung typgleicher Items

So bezieht sich beispielsweise die Bezeichnung B auf Items mit Brückeninformationen (sowohl Sachtext als auch narrativer Text und sowohl explizite als auch paraphrasierte Anforderungen), EL auf alle expliziten, lokalen Items (sowohl Sachtext als auch narrativer Text) und T auf alle temporale Items (sowohl Sachtext als auch narrativer Text und sowohl explizite als auch paraphrasierte als auch implizite Anforderungen). IKS4 bezeichnet beispielsweise ein implizites, kausales Item, dessen Textbasis ein Sachtext ist. Es ist das vierte Item seiner Art.

Aus messtheoretischer Sicht wäre es wünschenswert, wenn jede Kombination dieser drei Parameter mit mindestens drei Items in der vorläufigen Testendversion vertreten ist. Daher sollen mindestens drei Items pro Kombination erstellt werden, aus denen gegebenenfalls die geeignetsten Items ausgewählt werden sollen. Die Auswahl der Items für die vorläufige Testendversion erfolgt in zwei Stufen und ist in Kapitel 6 und in Kapitel 7 beschrieben. Die Anzahl des Itempools soll überdies aus Gründen der Ökonomie möglichst gering gehalten werden.

Auf Basis dieser Vorgaben wurden 108 Items erstellt, deren Verteilung auf die einzelnen Kombinationen Tabelle 12 entnommen werden können. Nicht alle von ihnen sollen in der vorläufigen Testendversion Anwendung finden.

Tabelle 12: Erstellte Testitems

	B		**L**		**T**		**K**		**Summe:**
	S	**N**	**S**	**N**	**S**	**N**	**S**	**N**	
E	4	5	4	4	4	4	4	4	**32**
P	3	4	4	4	4	4	4	4	**32**
I	-	-	10	4	10	4	12	2	**44**
Summe:	**7**	**9**	**18**	**12**	**18**	**12**	**20**	**12**	

5.4 Generierung der Testitems

Essenziell für die Umsetzung des Itemschemas ist die Sicherstellung der Eindimensionalität jener Testitems, die in ihrer Kombination aus Aufgaben-, Informations- und Textart identisch ist. Um diese Eindimensionalität zu gewährleisten, ist eine systematische, maximal standardisierte Itemgenese unabdingbar. Ihre Umsetzung soll nun skizziert werden.

Für jede mögliche Informationsart wurde zunächst ein prototypisches Item (*Dummy*) entwickelt. Dieses umfasste eine Textbasis, eine Frage, eine richtige Antwort und drei Distraktoren. Dieser Prototyp zielt jeweils auf einen spezifischen Aufgabentyp ab. Um ihn auf weitere Aufgabentypen zu übertragen, musste er jeweils leicht abgewandelt werden, beispielsweise indem die gesuchte Information im Aufgabenstamm paraphrasiert wurde. Überdies wurde für jeden Aufgabentyp eine eigene Rückmeldung erstellt, die für alle Items dieses Aufgabentyps gleich aufgebaut war. Damit lag für jede Kombination von Aufgaben- und Informationsart eine spezifische Vorlage vor. Anhand dieser Vorlagen wurden systematisch die Items generiert, die sich nur im Aufgabenstamm unterscheiden sollten. Somit sind sich die Items jeder einzelnen Facettenkombination in ihrer Struktur maximal ähnlich und zielen stets auf dieselbe Aufgaben- bzw. Informationsart ab, auch wenn sie unterschiedliche Situationsmodelle erfassen. Abschließend wurde für jede Informationsart ein *Dummy*item aus der Vorlage ausgewählt und dem Itempool hinzugefügt. Diese *Dummy*items haben die Bezeichnung PBN1, PLN1, ITN1 und EKN1. Damit ist je ein Item pro Vorlage im Itempool vorhanden. Es konnte pro Vorlage jeweils nur ein Item in den Itempool aufgenommen werden. Hätte man alle drei Items der Vorlage aufgenommen, so wären im Itempool Aufgaben, die sich hinsichtlich ihrer Textbasis nicht voneinander unterscheiden. Eine solche Konfundierung der Items steht jedoch im Widerspruch zu den in Kapitel 5.1 getroffenen Prämissen der Itemgenese.

Generierung der Feedbacks

Das erstellte Feedback besteht aus einer Rückmeldung, ob die vom Testanden gegebene Antwort richtig oder falsch ist. Bei einer falschen Antwort wird eine

Hilfestellung gegeben. Diese enthält eine Aufforderung, die von der Art der Aufgabe (lokalisieren, paraphrasieren oder integrieren) abhängig ist. Die Aufforderungen im Feedback werden von der Aufgabenart bestimmt und sind bei allen Informationsarten gleich. So werden beispielsweise Aufforderungen bei expliziten Items immer zum „Suchen nach der Stelle im Text“ auffordern, gleichgültig, ob die Items beispielsweise auf lokale oder auf kausale Informationen abzielen.

Bei den Rückmeldungen wird der Fokus des Lesers so auf die Hinweisreize im Aufgabenstamm gelenkt, die für die Lösung der Aufgabe relevant sind. Es wird beispielsweise auf lösungsimmanente Textstellen explizit aufmerksam gemacht und kausale Zusammenhänge oder zeitliche Abfolgen werden durch Signalwörter (z. B. wann) ins Blickfeld gerückt. Diese Feedbackart hat sich in Hinblick auf die Lesekompetenz als besonders effektiv erwiesen (Golke et al., 2015).

Aspekte der Sicherung der Inhaltsvalidität und der Prämisse der Itemunabhängigkeit bei gleichzeitiger Eindimensionalität

Um eine valide Umsetzung im Sinne der theoriegeleiteten Vorgaben sicherstellen zu können, ist es nötig, einige Gesichtspunkte im Besonderen zu berücksichtigen. So muss bei der Erstellung des Itempools auf Basis der *Dummy*items darauf geachtet werden, dass Aufgabenstamm, Fragestellung und Antwortalternativen nach Möglichkeit derart gestaltet sind, dass das Lösen der Aufgabe nicht ausschließlich an der Erkennung eines einzelnen Wortes liegt. In diesem Fall würde der Lesekompetenztest zu sehr auf Worterkennung abzielen und damit die eigentlich zu erfassende Leseleistung aus dem Fokus verlieren, was insbesondere bei impliziten Items problematisch wäre. Wenn beispielsweise die richtige Antwort eines lokalen Items „*unter der Decke*“ ist und ein Distraktor „*auf der Decke*“ lautet, dann ist die Fähigkeit, zwischen den Wörtern „*unter*“ und „*auf*“ zu diskriminieren essenziell zum Lösen der Aufgabe und erhält ein unangemessen starkes Gewicht in der durch dieses Item erfassten Lesekompetenzfacette. Wenn dieses Item nun implizit wäre, so würde es sich systematisch von anderen lokalen impliziten Items unterscheiden, bei denen das Ziehen von Inferenzen im Vordergrund steht. Damit würde es nicht

nur eine reduzierte Inhaltsvalidität aufweisen, sondern auch der Prämisse der Eindimensionalität aller typgleichen Items entgegenwirken.

Ein weiterer für die Itemgenese bedeutsamer Aspekt betrifft die Verschränkung der Informationsarten untereinander (vgl. Kapitel 3.4) sowie die Tatsache, dass in der Textbasis in der Regel Informationen zu mehreren Dimensionen gegeben sind (beispielsweise wird ein zeitlicher Verlauf beschrieben, in welchem die Entitäten sich von einem Ort zum anderen bewegen bzw. bewegt werden). Es muss explizit darauf geachtet werden, dass Frage und Antwortalternativen derart gestaltet sind, dass sie genau nur eine Dimension abfragen. Dies ist insbesondere bei kausalen Items von Bedeutung, bei denen durch Antezedenz und Konsequenz bereits ein zeitlicher Verlauf implizit vorhanden ist, welcher für den nicht temporalen Aufgabenfokus jedoch nicht von Interesse ist (siehe hierzu auch die Ausführungen zu den Dimensionen zweiter Ordnung in Kapitel 3.4).

Für die temporalen Items wird außerdem besonders darauf geachtet, dass vom Leser kein Umgang mit Zeiten und Zahlen erwartet wird, welcher bereits in den Kompetenzbereich Rechnen fällt. So soll gewährleistet werden, dass der Lesekompetenztest nur die Kompetenzdimension Lesen erfasst.

Insgesamt ist für die Itemgenese wichtig, die theoriegeleiteten Vorgaben möglichst systematisch zu implementieren und damit die Inhaltsvalidität des Testmaterials zu gewährleisten. Neben dieser sind für die Qualität des Testverfahrens jedoch noch weitere Aspekte von Bedeutung, die im Anschluss an die Materialentwicklung empirisch zu überprüfen sind.

6 Qualitative Vorerprobung des dynamischen Lesekompetenztests

6.1 Fragestellung

Das primäre Ziel der qualitativen Vorerprobung liegt darin, die Passung zwischen der Zielpopulation und den bislang entwickelten Materialien zu eruieren. Auf Grundlage dieser ersten Erhebung soll der zu konstruierende Test weiterentwickelt und optimiert werden. Wie bei qualitativen Untersuchungen üblich, gibt es keine konkreten Hypothesen, die getestet werden sollen. Vielmehr soll eine neue Erhebungsmethode erstmals explorativ erprobt werden. Mögliche Problembereiche der Items können so vor der Pilotierung erkannt und behoben werden.

Neben der allgemeinen Performanz liegt der Fokus dieser Untersuchung auf den während der Aufgabenbearbeitung ablaufenden Gedankengängen, die Rückschlüsse auf Probleme bei der Anwendung oder auf Interpretations- und Verständnisschwierigkeiten zulassen. Daneben sind einige im Fokus der Erhebung stehende Aspekte die allgemeine Verständlichkeit der Aufgaben und Feedbacks, das Interesse der Kinder an den Themen und Texten, die Motivation und eventuelle Ermüdungseffekte sowie mögliche unbekannte Wörter, die Responsivität auf das gegebene Feedback und insbesondere auch die Schwierigkeit der einzelnen Aufgabe und deren mögliche Quellen.

6.2 Methodik

6.2.1 Stichprobe

Die Untersuchung wurde an insgesamt 15 Schülern am Ende der dritten Jahrgangsstufe durchgeführt. Die Schüler kamen aus zwei baden-württembergischen Grundschulen.

Aus Schule A wurden insgesamt 13 Schüler der dritten Klasse untersucht. Ein Kind überschritt das a priori angesetzte Zeitlimit der Testung von 45 Minuten und musste durch einen weiteren Probanden ersetzt werden. Das durchschnittliche Alter der Probanden lag bei 9;3 Jahren (SD=0;3). 41.6 % der Kinder waren weiblich. Fünf Erziehungsberechtigte von Schule B gaben ihr Einverständnis zur Untersuchung, davon konnten drei Kinder valide erhoben werden. Das durchschnittliche Alter der Probanden liegt bei etwa 9 Jahren, exakte Angaben zu Alter und Geschlecht der Probanden liegen für diese Schule nicht vor. Zwei Kinder gingen fälschlicherweise davon aus, dass es sich um eine Testung mit Relevanz für die Notengebung handelte. Alle Kinder stammen aus der dritten Jahrgangsstufe (vgl. Autenrieth, 2014; Weidner, 2014). Die Zuteilung der Versuchspersonen zu den Untersuchungsbedingungen erfolgte stets randomisiert.

6.2.2 Design und Ablauf der Erhebung

Methode des lauten Denkens

Kennzeichnend für die in dieser Erhebung zum Einsatz kommende qualitative Methode des lauten Denkens ist die Erfassung zusätzlicher verbaler Information neben der eigentlichen Testperformanz. Diese erfolgt, indem der Proband nicht nur die ihm gestellte Frage beantwortet bzw. die ihm gestellte Primäraufgabe löst, sondern dabei auch seine Gedankengänge verbalisiert (*think aloud*). Auf diese Weise sollen die ablaufenden gedanklichen Prozesse des Probanden erfasst werden (Völzke, 2012). Bei der hier beschriebenen Erhebung wurde die Methode des lauten Denkens (*think aloud*) mit den sogenannten *follow-up*-Nachfragen kombiniert, bei denen sich die bislang

gewonnene Information anschließend durch strukturiertes Nachfragen gezielt erweitern und ergänzen lässt (Rubin & Rubin, 2011). Im Bereich der Leseforschung haben sich *think-aloud*-Protokolle bei Kindern bereits als valide Messinstrumente bewährt (Coté & Goldman, 1999).

Untersuchungsdesign

Zur Umsetzung der qualitativen Methodik wurde ein Versuchsleiter-Leitfaden mit standardisierten Anweisungen und *follow-up*-Fragen entwickelt, um ein gewisses Maß an Standardisierung gewährleisten zu können. Daneben wurde in Anlehnung an das lateinische Quadrat (Rack & Christophersen, 2009, S. 22; Toutenbourg, 1994, S. 159) ein Untersuchungsdesign entwickelt, welches die zu testenden 108 Items auf 12 Gruppen à 9 Items verteilt. Jeder Proband wurde einer Gruppe zugewiesen. Insgesamt sollte die Testprozedur für jeden Probanden eine Schulstunde nicht überschreiten. Die Items wurden derart verteilt und in ihrer Reihenfolge determiniert, dass die geschätzte Aufgabenschwierigkeiten innerhalb einer Sitzung ansteigen und sich in ihrer Summe zwischen den Untersuchungsgruppen nicht wesentlich unterscheiden sollten. Die Aufgabenschwierigkeit wurde dabei in Anlehnung an das in Kapitel 5.2 erwähnte Schema von Kirsch (2001) abgeschätzt.

Um die Kinder mit der Methode des lauten Denkens vertraut zu machen, wurden den eigentlichen Aufgaben drei Probeitems vorgeschaltet. Als Probeitems wurden die *Dummy*items verwendet, die nicht in den eigentlichen Aufgabenpool Eingang fanden. Bei der Zuteilung der Probeitems auf die Versuchsgruppen wurde berücksichtigt, dass unter den eigentlichen Items auch solche mit zu den Probeitems identischem Aufgabenstamm vorliegen können. In diesem Fall musste der Permutationsplan dieser Doppelung Rechnung tragen. Bei der Versuchsplanerstellung wurde außerdem darauf geachtet, dass sich die Probeitems und die eigentlichen Aufgaben in ihrer Kombination aus Aufgaben- und Informationsart hinreichend unterscheiden und keine Übungseffekte auftreten konnten. Gegebenenfalls wurde das Design so adaptiert, dass jeder Aufgabenstamm nur einmal pro Gruppe vorkam. Der vollständige Permutationsplan ist in Anhang B zu finden.

Testleiterinnen

Die Erhebung wurde von zwei Studentinnen der Pädagogischen Hochschule Heidelberg im Rahmen ihrer wissenschaftlichen Hausarbeiten durchgeführt (Autenrieth, 2014; Weidner, 2014). Neben der schriftlichen Anweisung in Form des Leitfadens erfolgte eine umfassende mündliche Instruktion der Testleiterinnen, da qualitative Untersuchungen stark von der Person des Testleiters und ihrer Wechselwirkungen mit dem Probanden beeinflusst werden.

Ablauf der Erhebung

Die qualitative Untersuchung wurde als Einzelsitzung von etwa einer Stunde Dauer durchgeführt und fand gegen Ende des Schuljahres 2013/14 statt.

Der Ablauf erfolgte standardisiert: Zunächst wurde eine kleine Fantasieübung durchgeführt, die neben der Vorbereitung auf die eigentliche Testung auch eine „Eisbrecherfunktion“ hatte. Anschließend bearbeitete das Kind die drei Probeitems, auf die die eigentlichen Aufgaben in der durch die Versuchsgruppe determinierten Reihenfolge folgten. Je nach Performanz des Kindes bei den Probeitems wurde teilweise bereits nach einem oder zwei Probeitems mit der eigentlichen Erhebung begonnen. Jedes Item wurde als Aufgabenstamm mit Frage und Antwortalternativen auf Papier gedruckt dargeboten und vom Probanden gelesen. Die Position der richtigen Antwort unter den vier Antwortalternativen wurde für jedes Item zufällig bestimmt, hierfür wurden mit der Software *R* (R Core Team, 2014) Zufallszahlen zwischen 1 und 4 erzeugt. Das Kind verbalisierte seine Überlegungen und seine Antwort. Bei einer falschen Antwort erhielt es von der Testleiterin das ebenfalls auf Papier gedruckte Feedback. Der Proband las das Feedback und versuchte, die Aufgabe nochmals zu lösen. Nach diesem zweiten Versuch oder auch nach einer richtigen Antwort im ersten Versuch folgte das nächste Item. Nach jeder Aufgabe stellte die Testleiterin gezielte Nachfragen (*follow-up*). Die Untersuchung endete mit allgemeinen Nachfragen zur gesamten Erhebung. Die Themenbereiche, die sowohl durch die gezielten als auch durch die allgemeinen Nachfragen abgedeckt werden sollten, waren vorab durch den Versuchsleiterleitfaden spezifiziert und den Testleiterinnen an die Hand

gegeben worden. Jede Sitzung wurde für die nachfolgende Datenauswertung mit Tonband aufgezeichnet, die Testleiterin machte ergänzende Notizen, die sich auf den Probanden (beispielsweise in Bezug auf Temperament oder Müdigkeit) und die Untersuchungssituation (beispielsweise die allgemeine Atmosphäre während der Testung oder eventuelle externe Störungen) bezogen.

6.2.3 Datenauswertung

Die Auswertung der auf diese Weise gewonnenen Informationen erfolgte in drei Schritten. Zunächst wurden die Sitzungen transkripiert. Anschließend wurden die Transkripte und alle weiteren vorliegenden Informationen zu einem Exzerpt aggregiert. Schließlich konnte an Hand der Exzerpte eine umfassende Analyse der vorliegenden Informationen durchgeführt werden.

Für die Transkription wurde in Anlehnung an etablierte Transkriptionsregeln (z. B. Selting et al., 2009; Göpferich, 2007) eine standardisierte Anleitung und eine Vorlage entwickelt, die durchgehend verwendet wurde. Der Fokus lag hierbei auf dem Lösungsweg der einzelnen Aufgaben. Transkriptstellen, die Rückschlüsse auf Itemschwierigkeiten oder Begründungen für gewählte Antworten zuließen, wurden entsprechend markiert. Die Zeitspannen, die ein Kind zur Lösung einer Aufgabe benötigte, flossen ebenfalls in das Transkript ein. Die Transkriptvorlage und die Anleitung zum Transkripieren können bei der Autorin angefordert werden.

Um ein Maximum an Information aus den Transkripten abzuschöpfen, wurden diese zu Exzerpten verdichtet, wobei auch die Notizen und verbalen Berichte der Testleiterinnen miteinbezogen wurden. Dazu wurden die Transkripte zunächst gesichtet. An Hand dieses ersten Eindrucks wurden Leitfragen entwickelt, die für ein bestimmtes Item die Informationen erfragten, welche essentiell für die Weiterentwicklung des Testmaterials waren. Diese Leitfragen zielten auf Motivation, Interesse und Performanz der Kinder ab sowie auf Verständnisprobleme bei der Aufgabe (z. B. durch unbekannte Wörter) und auf die Verständlichkeit und die Wirkung des gegebenen Feedbacks. Auch wurde nach den jeweiligen Gedankengängen beim Bearbeiten der Aufgabe, nach

eventuell vorhandenem Vorwissen und dessen Auswirkungen auf den Bearbeitungsprozess gefragt. Weitere Themenfelder waren Probleme des Probanden mit der Methode des lauten Denkens, die allgemeine Testsituation und die Beziehung zwischen dem Probanden und der Testleiterin. Diese Leitfragen wurden in einer Exzerptvorlage zusammengestellt, welche für jedes Item standardisiert auszufüllen war.

Da die Testleiterinnen explizit Gegenstand des Exzerpts waren, konnten sie bei der Erstellung der Exzerpte nicht eingesetzt werden, ohne die Auswertungsobjektivität in Frage zu stellen. Daher erstellten zwei studentische Mitarbeiterinnen nach einer umfassenden Schulung an Hand der Leitfragen für jedes Item und jeden Probanden jeweils ein Exzerpt. Für diesen Auswertungsschritt wurden zwei voneinander unabhängige Mitarbeiter eingesetzt, um Effekte zu minimieren, die in der Person des Exzerpterstellers begründet sind.

Damit lagen für jedes getestete Item vier Exzerpte vor, wenn zwei Kinder dieses Item bearbeitet hatten. Wurde ein Item nur von einem Kind bearbeitet, so lagen für dieses Item zwei Exzerpte vor. Insgesamt wurden für die getesteten 81 Items 270 Exzerpte erstellt.

Für jedes Item wurden alle Exzerpte zusammenfassend gesichtet und unter Berücksichtigung aller vorhandenen Informationen aggregiert. Um eine möglichst große Auswertungsobjektivität zu gewährleisten wurde dieser Schritt nicht von den Exzerpt- und Transkripterstellerinnen sondern von der Autorin vorgenommen. Die so extrahierten Informationen werden in Kapitel 6.3 vorgestellt.

Um die Schwierigkeit der Items besser quantifizieren zu können, wurde darüber hinaus für jedes Item unter Berücksichtigung aller verfügbarer Informationen seine Schwierigkeit auf einer Skala von 1 (sehr einfach) bis 5 (sehr schwer) eingeschätzt. Dabei wurden objektivere Schwierigkeitsparameter stärker gewichtet als subjektive Schwierigkeitsparameter. Zu den objektiven Schwierigkeitsparametern zählte beispielsweise, wie viele Kinder die Aufgabe

auf Anhieb richtig lösen konnten. Zu den subjektiven Schwierigkeitsparametern gehörte unter anderem die subjektive Einschätzung der Schwierigkeit durch die Kinder. Auch die Beobachtungen der Testleiterinnen flossen in den Schwierigkeitsindex mit ein. Die Selbstauskunft der Kinder wurde zunächst in einen vorläufigen Schwierigkeitsindex zwischen 1 und 5 übersetzt, der nachfolgender Tabelle entnommen werden kann. Analog wurde auch der objektive Schwierigkeitsparameter in einen Zahlenwert zwischen 1 und 5 transformiert.

Tabelle 13: Übersetzung der selbstberichteten Schwierigkeit eines Items in einen Schwierigkeitsindex

Selbsteingeschätzter Schwierigkeitsindex	**Selbsteinschätzung des Probanden**
1	leicht, einfach
2	relativ leicht, relativ einfach, bisschen leicht, bisschen einfach
3	nicht ganz so leicht, nicht allzu leicht, mittelschwer
4	relativ schwierig, relativ schwer, eher schwierig, eher schwer
5	schwierig, schwer

Der selbstberichtete Schwierigkeitsindex wurde bei Bedarf nach oben oder unten korrigiert, je nachdem, ob das Kind das Item lösen konnte und ob zur Lösung des Items ein oder zwei Versuchte benötigt wurden. Wurde beispielsweise ein Item sofort gelöst und enthielt es keine schwierigen oder unbekannten Wörter und wurde es auch vom Kind als einfach eingeschätzt, so wurde seine Schwierigkeit auf 1 gesetzt. Fand ein Kind ein Item schwierig und konnte es trotz Feedback nicht lösen, so wurde die Itemschwierigkeit für dieses Item auf 5 gesetzt. Ein selbsteingeschätzter Schwierigkeitsindex von 3 wurde beispielsweise auf 4 erhöht, wenn das Item nur mit Hilfe des Feedbacks lösbar war. Wenn das Item relativ mühelos ohne Zuhilfenahme des Feedbacks lösbar

war, der selbsteingeschätzte Schwierigkeitsindex jedoch bei 5 („schwierig“ bzw. „schwer“) lag, so wurde der finale Schwierigkeitsindex von 5 auf 3 verringert, da die subjektive Selbsteinschätzung gegenüber der tatsächlichen Performanz als weniger stark zu gewichten war.

Wurde ein Item von zwei Probanden bearbeitet, so wurden beide Probanden für die finale Indexbildung berücksichtigt. Differierten die Schwierigkeitsindizes der Probanden (beispielsweise einmal 1 und einmal 5), so wurde mit Blick auf eine spätere Anwendung bei Kindern mit spezifischem Förderbedarf besonders der höhere Schwierigkeitsindex für die Bildung des finalen Index berücksichtigt. Dieser finale Schwierigkeitsindex des Items wurde in Bezug zu den Itemausprägungen Informationsart, Aufgabenart und Textart gesetzt.

6.3 Ergebnisse

Insgesamt liegen zu 81 Items 135 Antworten von 15 Kindern vor. Bei einem Probanden musste die Testung abgebrochen werden. Die Testbedingung wurde mit einem anderen Kind wiederholt (Autenrieth, 2014). Gemäß eigenen Angaben finden die meisten Kinder die Aufgaben leicht, das Interesse an den in den Aufgaben enthaltenen Themen ist meist gegeben. Die Kinder waren laut Beobachtungen der Testleiterinnen bis auf eine Ausnahme motiviert und in ihrer Leistungsfähigkeit nicht durch Ängstlichkeit eingeschränkt (Autenrieth, 2014; Weidner, 2014). Hinweise auf problematische Beziehungen zwischen den Kindern und den Testleiterinnen liegen nicht vor. Ermüdungserscheinungen auf Seiten der Probanden zeigen sich kaum, wenn es zu Müdigkeit kam, dann immer gegen Ende der Testung. Insgesamt finden sich auch bei ein- und demselben Item teilweise sehr große Unterschiede zwischen den Kindern, ein Item, das dem ersten Kind sehr leicht fällt, kann für das zweite Kind sehr anspruchsvoll sein. Größere Verständnisschwierigkeiten bleiben generell aus, gelegentlich sind einige Wörter nicht bekannt. Diese sind nachfolgender Tabelle zu entnehmen.

Tabelle 14: Verständnisschwierigkeiten in der qualitativen Vorerprobung

Item	**Unbekannte Wörter**	**Wie vielen Kindern unbekannt?**
EBS2	Wirkung	1 von 2
ELS4	Zentralamerika	1 von 1
PLN2	Bienenstöcke	1 von 2
ILS1	Gelege	1 von 2
ETS2	Schlacke, Roheisen	1 von 2
ITS4	Essgewohnheiten	1 von 2
IKS5	Fortpflanzen	1 von 1
IKS7	Kohlenstoff	1 von 1
IKS9	Graphit(teilchen)	2 von 2
IKS10	Wasserstoff	1 von 2

Zum Feedback

Das Feedback wird von den Kindern meist nicht in Anspruch genommen. Insgesamt kommt es in 30 Fällen zum Einsatz. Wie aus Tabelle 15 hervorgeht,

kann die Aufgabe nach dem Feedback häufig korrekt gelöst werden. In der Regel wird das Feedback von den Probanden als hilfreich erlebt. Ist dies nicht der Fall, so lassen sich dafür meist vier Gründe finden. Zum einen verunsichert das Feedback manchmal, es wird als zusätzliche Aufgabe wahrgenommen. Zum anderen liegt der Fokus des Probanden trotz der Rückmeldung weiterhin auf der Anwendung einer falschen Strategie oder auf nicht lösungsimmanenten Textteilen. Beispielsweise wird das Kind bei einem impliziten Item zum Überlegen/Ziehen einer Inferenz ermuntert, das Kind sucht jedoch weiter nach der expliziten Lösung im Aufgabenstamm. Ein dritter Grund, der nach den Befunden der qualitativen Vorerprobung eine Wirksamkeit des Feedbacks verhindert, ist das Vorwissen der Kinder, wenn es in einem Spannungsverhältnis zum Aufgabenstamm steht. Hierbei scheint das Vorwissen die im Textstamm enthaltene Information stets zu überlagern. Daneben können einzelne, unbekannte Wörter verhindern, dass das Feedback richtig verstanden und umgesetzt wird.

Tabelle 15: Feedback in der qualitativen Vorerprobung: Beanspruchung und Responsivität

	relative Häufigkeit	**absolute Häufigkeit**
Feedback in Anspruch genommen	22.22 %	30
Aufgabe nach Feedbackdarbietung gelöst	15.56 %	21
Aufgabe nach Feedbackdarbietung nicht gelöst	6.66 %	09

Zur Itemschwierigkeit

Die Schwierigkeit der Items wurde in einem Schwierigkeitsindex von 1 (sehr einfach) bis 5 (sehr schwer) abgebildet und in Bezug zu den Itemausprägungen Informationsart, Aufgabenart und Textart gesetzt. Abbildungen 9 bis 11 machen die Zusammenhänge deutlich. Auf Grund des geringen Stichprobenumfangs wurde von einer statistischen Absicherung abgesehen. Stattdessen soll eine rein deskriptive Datenbewertung erfolgen.

Eine grafische Inspektion der Abbildung 9 zeigt die Schwierigkeit verschiedener Itemgruppen mit unterschiedlicher Informationsart und Textart.

Die Schwierigkeit ist auf der y-Achse abgetragen. Außer bei den Brückeninformationen scheinen Sachtexte leicht schwieriger zu sein. Der Effekt der Informationsart auf die Schwierigkeit scheint sich zwischen Sachtexten und narrativen Texten kaum zu unterscheiden. Temporale Items scheinen generell mit einer leicht erhöhten Schwierigkeit assoziiert zu sein, jedoch ist der Schwierigkeitsindex nicht größer als der Mittelpunkt der fünfstufigen Schwierigkeitsskala. Auffallend ist, dass kausale Items nicht schwieriger zu sein scheinen als lokale Items. Über beide Textarten hinweg kann der Abstand zwischen Items mit Brückeninformationen und Items mit temporalen Informationen als besonders groß angesehen werden.

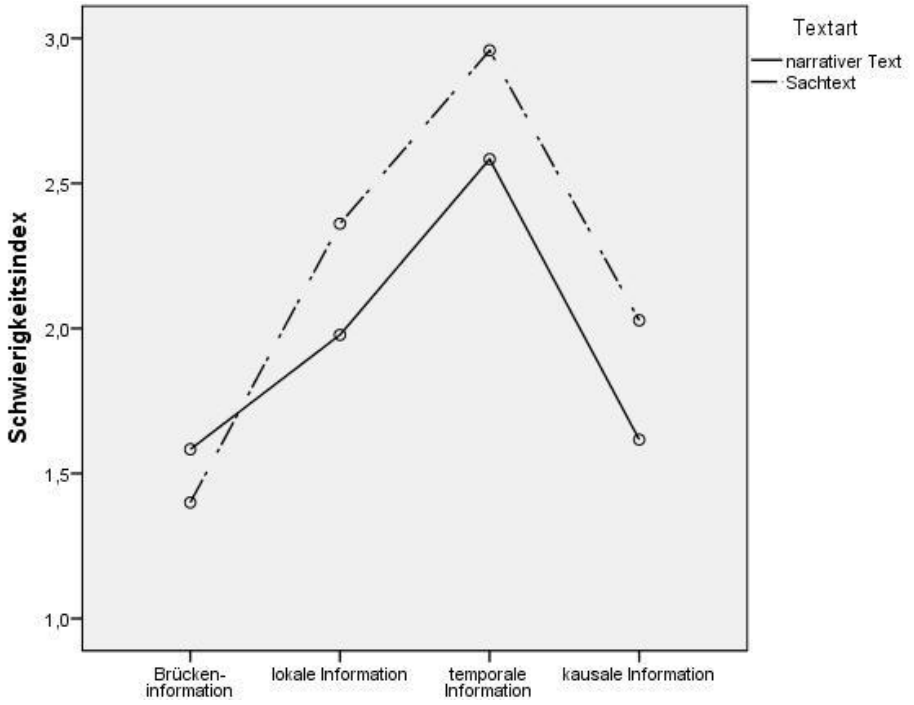

Abbildung 9: Qualitative Vorerprobung: Auswirkungen der Informationsart auf die Schwierigkeit bei unterschiedlichen Textarten

In Abbildung 10 ist die Schwierigkeit verschiedener Itemgruppen mit unterschiedlicher Aufgabenart und Textart dargestellt. Die Schwierigkeit ist wiederum auf der y-Achse abgetragen. Außer bei den expliziten Items scheinen Sachtexte leicht schwieriger zu sein. Das Muster der Schwierigkeit scheint sich zwischen Sachtexten und narrativen Texten nicht zu unterscheiden. Auffallend ist die erhöhte Schwierigkeit bei impliziten Aufgaben, während sich explizite

und paraphrasierte Items nicht zu unterscheiden scheinen. Sachtexte scheinen in der impliziten Bedingung besonders anspruchsvoll zu sein.

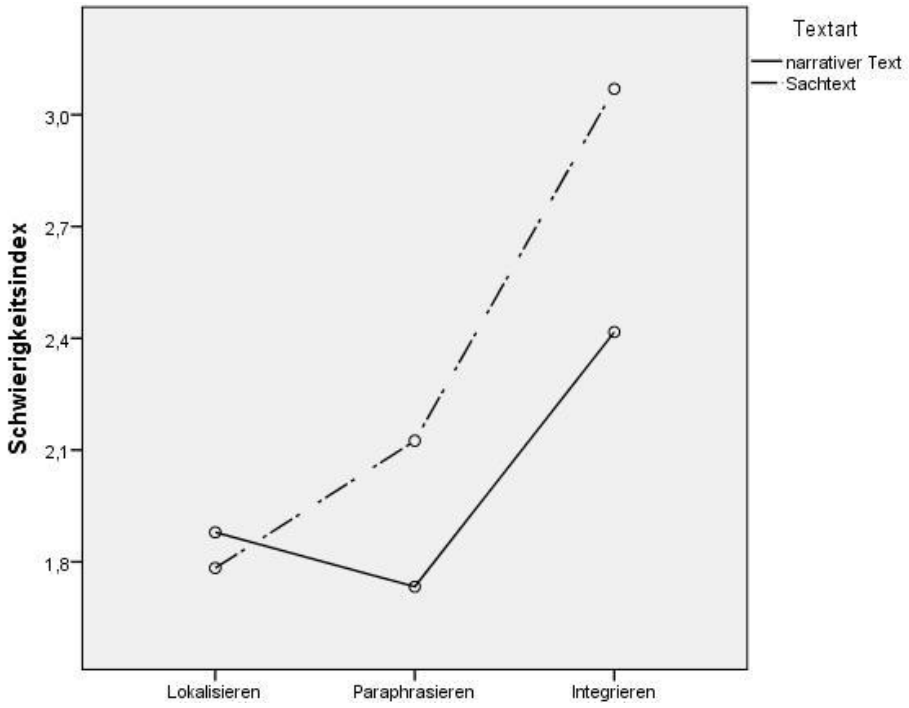

Abbildung 10: Qualitative Vorerprobung: Auswirkungen der Aufgabenart auf die Schwierigkeit bei unterschiedlichen Textarten

Der Kombination von Informations- und Aufgabenart kommt eine wichtige Bedeutung zu, da sie die eigentlichen Facetten des zu entwickelnden Lesekompetenztests darstellen. Aus diesem Grund sollen sie nochmals gesondert dahingehend betrachtet und auf Interaktionseffekte inspiziert werden. Da zwischen narrativen Texten und Sachtexten bislang keine übermäßigen Unterschiede erkannt wurden, sollen sie bei der nachfolgenden Betrachtung vernachlässigt werden.

Die Schwierigkeit verschiedener Itemgruppen mit unterschiedlicher Informationsart und Aufgabenart kann Abbildung 11 entnommen werden. Die Schwierigkeit ist wieder auf der y-Achse abgetragen. Der bislang entstandene Eindruck über die Determinanten der Schwierigkeit bestätigt sich hier. Implizite Items sind besonders anspruchsvoll, unabhängig davon, welche Informationsart erfragt wird. Temporale Items scheinen bei allen

Aufgabenarten besonders anspruchsvoll zu sein. Insgesamt zeigen bezüglich der lokalen, temporalen und kausalen Informationsarten explizite Aufgaben den minimalsten Schwierigkeitsindex.

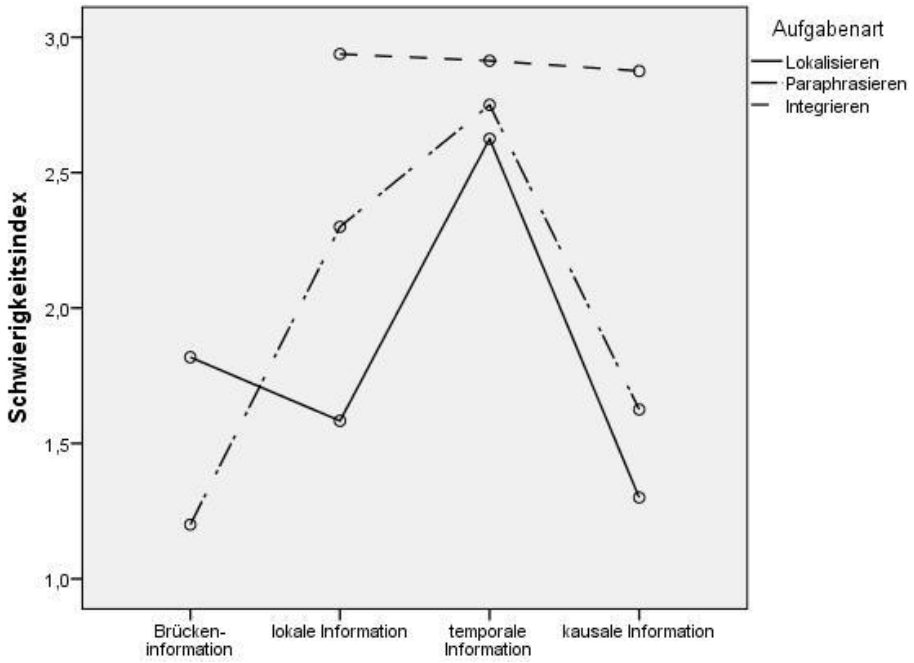

Abbildung 11: Qualitative Vorerprobung: Auswirkungen der Informationsart auf die Schwierigkeit bei unterschiedlichen Aufgabenarten

6.4 Interpretation und Implikation für die weitere Testentwicklung

Auf Grundlage der qualitativen Vorerprobung können die bislang entwickelten Testmaterialien weiterentwickelt werden. Es werden etwa ein Viertel der überprüften Items überarbeitet und optimiert, meist werden dabei unbekannte Wörter durch einfachere Ausdrücke ersetzt (vgl. Tabelle 14) und Situationsmodelle vereinfacht.

Generell scheinen die Kinder das Testmaterial gut angenommen zu haben, was sich sowohl in den als einfach empfundenen Aufgaben, als auch in der hohen Anzahl richtig gelöster Items zeigt. Deckeneffekte sind nicht ganz auszuschließen, jedoch in Hinblick auf das von den Testleiterinnen berichtete generell erhöhte Leistungsniveau der getesteten Kinder einerseits (Autenrieth, 2014) und der Zielpopulation mit besonderem Förderbedarf andererseits als eher unkritisch zu bewerten. Interesse und Motivation scheinen hinreichend gut gegeben zu sein.

Theoriekonform ist das Ziehen von Inferenzen etwas anspruchsvoller als das Lokalisieren und Paraphrasieren, während explizite Items bei den im Testfokus stehenden Informationsarten die geringste Schwierigkeit aufweisen. Überraschend ist jedoch die relativ zu den anderen Bedingungen erhöhte Schwierigkeit bei temporalen Aufgaben, welche auch erklären könnte, warum sich bei temporalen Items die Schwierigkeitsindizes zwischen den einzelnen Aufgabenarten relativ gering voneinander unterscheiden. Dies kann dem höheren Abstraktionsniveau temporaler Items geschuldet sein, da zeitliche Abläufe per se nicht sinnlich erfasst werden können. Nicht ganz auszuschließen ist jedoch auch, dass es sich bei dieser Auffälligkeit um eine stichprobenspezifische Besonderheit handelt, welche in anderen Populationen nicht in dieser Form zu tragen kommt. Spezifika des einzelnen Probanden fallen bei dieser Untersuchung besonders ins Gewicht, was neben der geringen Probandenanzahl, die jeweils ein Item bearbeitet hat auch der qualitativen Methodik dieser Untersuchung geschuldet ist. Wenn es sich bei dem hier beobachtbaren Muster um eine spezifische Interaktion zwischen den Items und genau dieser Stichprobe handelt, so sollten sich diese Muster im nächsten

Schritt der Testkonstruktion, der Pilotierung, in dieser Form nicht mehr finden lassen.

7 Pilotierung des dynamischen Lesekompetenztests

7.1 Fragestellung

Ziel der Pilotierung ist die Erhebung wichtiger teststatistischer Kennwerte an einer umfassenden Stichprobe. Auf Basis dieser sollen Items für die vorläufige Testendversion ausgewählt und diese erstellt werden. Neben dem Gesamttest sind dabei auch die einzelnen Subskalen Aufgabenart und Informationsart von Interesse. Spezifisch sollen folgende Fragestellungen auf der Analyseebene der Skalen untersucht werden:

P.1. Dimensionalitätsprüfung

P.1.1. Welche Reliabilitäten weisen die einzelnen Subskalen des Tests und der Test als Ganzes auf?

Um die Reliabilitäten angemessen interpretieren zu können, muss die Eindimensionalität der Skala zwingend gewährleistet sein. Daraus leitet sich eine weitere Fragestellung ab:

P.1.2. Sind die einzelnen Subskalen des Tests eindimensional und ist der Test als Ganzes eindimensional?

Auf Ebene der Items sind folgende Fragestellungen von Relevanz:

P.2. Reliabilitätsprüfung

P.2.1. Welche Itemschwierigkeiten weisen die Items der einzelnen Subskalen des Tests auf?

P.2.2. Welche Itemvarianzen weisen die Items der einzelnen Subskalen des Tests auf?

P.2.3. Welche Trennschärfen weisen die Items der einzelnen Subskalen des Tests auf?

P.2.4. Welche Selektionskennwerte weisen die Items der einzelnen Subskalen des Tests auf?

7.2 Methodik

7.2.1 Stichprobe

Die Pilotierungsstichprobe bestand aus 240 Schülern der dritten und vierten Jahrgangsstufe aus sechs Grundschulen aus der Region Rhein-Neckar. An keiner dieser Schulen wurde bereits im Rahmen der qualitativen Voruntersuchung erhoben. 124 Schüler (51.7 %) waren aus der dritten Jahrgangsstufe und 116 Schüler (48.3 %) aus der vierten Jahrgansstufe. Die Stichprobe umfasste 119 Mädchen (49.6 %) und 121 Jungen (50.4 %). Das durchschnittliche Alter betrug 9 Jahre und 3 Monate (SD=1;4 Jahre).

7.2.2 Design und Ablauf der Erhebung

Die Untersuchung basierte auf einem dreifaktoriellen, *within-subjects* Design. Der Faktor „Informationsart“ war vierstufig (Brückeninformationen, lokale, temporale und kausale Informationen), der Faktor „Aufgabenart“ umfasst drei Ausprägungen (explizite, paraphrasierte und implizite Aufgaben). Der Faktor „Textart“ war zweistufig (narrative Texte und Sachtexte). Da Brückeninformationen nicht implizit abgefragt werden können (vgl. Kapitel 5), ist das Design unvollständig.

Permutationsplan

Um die zur Beantwortung der Fragestellungen nötigen Berechnungen durchführen zu können, mussten besondere Anforderungen an den Permutationsplan gestellt werden. Jedes der 108 Items muss in der Pilotierung mit jedem anderen Item verlinkt sein, also von einer bestimmten Anzahl an Probanden in derselben Testung bearbeitet werden. Daneben gilt es, die Arbeitsbelastung für die Probanden zu begrenzen und aus Gründen der Ökonomie die Anzahl der Probanden möglichst gering zu halten. Diese Vorgaben stehen in einem Spannungsverhältnis zueinander, welches durch ein geeignetes Studiendesign aufgelöst werden muss.

In der Pilotierung wurden die zu evaluierenden 108 Items auf 56 Bedingungen aufgeteilt und jeder Proband wurde randomisiert einer Bedingung zugewiesen, wobei eine Ausbalancierung hinsichtlich Klassenstufe und Geschlecht anvisiert wurde. Die Bedingungen umfassten jeweils zwischen 24 und 30 Items und waren in ihrer Zusammensetzung so gewählt, dass die Arbeitsbelastung der einzelnen Bedingungen ähnlich war. Nachfolgend soll die Herleitung dieses Designs dargelegt werden.

Ausgangspunkt für die Erstellung eines geeigneten Permutationsplans war eine Modifizierung des Youden-Quadrats (Bortz, Lienert & Boehnke, 2008). Tabelle 16 gibt einen Überblick über die Anzahl der aufzuteilenden Items pro Aufgaben- und Inferenzkombination und gibt in Klammern die Zuordnung der jeweiligen Items zu einer von vier Gruppen (A-D) an. Jeweils acht typgleiche Items wurden derselben Gruppe zugeordnet. Es bleiben sechs implizit-lokale Items übrig, die jeweils zur Hälfte Gruppe C und Gruppe D zugeordnet wurden und sechs implizit-temporale Items, die jeweils zu Hälfte Gruppe A und Gruppe B zugeordnet wurden.

Tabelle 16: Übersicht über die Items und ihre Aufteilung in der Pilotierung

	Brücken-information	**Lokale Information**	**Temporale Information**	**Kausale Information**
Explizit	8 (C)	8 (A)	8 (D)	8 (B)
Paraphrasiert	8 (B)	8 (D)	8 (A)	8 (C)
Implizit	-	14 (C & D)	14 (B & A)	16 (A & D)

Diese Zuteilung der Restitems beruhte auf zwei Überlegungen. Zunächst war auf Grundlage der qualitativen Vorerprobung anzunehmen, dass temporale Items die höchste Schwierigkeit aufweisen. Zusätzliche Items, die implizit temporale Information beinhalten, sollten daher nach Möglichkeit der Gruppe zugewiesen werden, die bislang eher einfachere Items bearbeitet hat, also Gruppe B, nicht Gruppe C. Daneben war die Überlegung, dass zur selben

Informationsart innerhalb einer Gruppe nicht paraphrasieren und gleichzeitig auch integrieren als Anforderungen gestellt sollte. Eventuelle Übungs- oder auch *spill-over*-Effekte innerhalb einer Gruppe lassen sich zwar nicht ganz vermeiden, jedoch sollten sie weniger stark ausgeprägt sein, wenn innerhalb der gleichen Inferenz statt den beiden Anforderungen paraphrasieren und integrieren die beiden Anforderungen lokalisieren und integrieren gegeben sind. Implizite Items mit lokaler Information wurden daher D zugewiesen, implizite Items mit temporaler Information A.

Nach einer weiteren Prämisse des Permutationsplans sollen sich Anforderungen und Inferenzen für zwei direkt aufeinander folgende Items jeweils unterscheiden, um *spill-over*-Effekte (z. B. Strategie des Vorgängeritems funktioniert auch bei Nachfolgeritem, dieses wird dadurch als weniger anspruchsvoll erlebt) möglichst minimal zu halten. Ein weiterer Vorteil ist primär motivationaler Natur: Die Kinder bearbeiten nicht viele schwere Items (z. B. acht implizit-temporale Items) direkt hintereinander. Vielmehr werden dazwischen immer wieder einfachere Items präsentiert, die die Motivation der Kinder erhalten sollen.

Jede Gruppe wurde in acht Blöcke à drei Items aufgeteilt, die sich durch Abwechslung in Inferenz- und Aufgabenart auszeichnen. Pro Block ist die Aufgabenreihenfolge immer gleich determiniert: lokalisieren, paraphrasieren, integrieren. Auf Grund der drei zusätzlichen impliziten Restitems bestanden drei Blöcke aus vier Items und hatten die fest determinierte Reihenfolge lokalisieren, integrieren, paraphrasieren, integrieren. Die beiden impliziten Items sollten hierbei unterschiedliche Textarten aufweisen.

Nach Möglichkeit sollte die Textart innerhalb eines Blocks abgewechselt werden. Wegen der Restitems konnte in der Gruppe A bzw. D beim Übergang zwischen zwei Blöcken auf ein implizites, lokales Item direkt ein explizites, lokales Item bzw. auf ein explizites, temporales Item direkt ein implizites, temporales Item folgen. In diesem Fall wurde darauf geachtet, dass sich die aufeinander folgenden Items in ihrer Textart unbedingt unterschieden.

Die Gruppen wurden in einem nächsten Schritt jeweils halbiert, so dass aus jeder Gruppe zwei Untergruppen mit jeweils vier Blöcken hervorgingen. Die Untergruppen umfassten mindestens 12 und maximal 15 Items. Es bestand außerdem die Vermutung, dass sich Übungseffekte während der Testung nicht ganz ausschließen lassen. Dies sollte sich insbesondere auf die Items auswirken, die immer am Ende einer (Unter-)Gruppe präsentiert werden sollen (z. B. Items aus dem jeweils letzten Block). Daher sollten nur die Hälfte der Kinder die Untergruppen in aufsteigender Blockreihenfolge bearbeiten. Die andere Hälfte bekam die Blöcke in absteigender Reihenfolge präsentiert.

Somit wurden aus den ursprünglichen vier Gruppen (A-D) acht Untergruppen à vier Blöcke (A1, A2, B1, B2, etc.) abgeleitet, die jeweils in zwei Versionen vorlagen: aufsteigende und absteigende Blockreihenfolge. Insgesamt waren die Items damit in 16 Einheiten aufgeteilt. In Anhang C findet sich eine Übersicht über die Blöcke, Untergruppen und Gruppen und ihre Items. Die 16 Einheiten wurden nun miteinander kombiniert. Beispielsweise wurde die Untergruppe A1 jeweils mit den sieben Einheiten A2 bis D2 kombiniert, A2 jeweils mit den verbleibenden sechs Einheiten B1 bis D2 usw., sodass 28 Untergruppenkombinationen vorlagen. Damit wurde gewährleistet, dass jedes Item mit jedem anderen Item zusammen getestet werden kann. Diese 28 Kombinationen gab es sowohl für die Untergruppen mit aufsteigender Blockreihenfolge als auch für die Untergruppen mit absteigender Blockreihenfolge. Beispielsweise gab es die Kombination A1C2 und die Kombination C2A1, wobei die Untergruppen in A1C2 jeweils aufsteigende Blockreihenfolge hatten, die Untergruppen in C2A1 jeweils absteigende Blockreihenfolge. Somit war beispielsweise ELN1 als erstes Item in Block 1 ein Item, das in A1C2 am Anfang und in C2A1 am Ende der Untersuchung bearbeitet werden sollte. Damit ergeben sich insgesamt zweimal 28 Kombinationen, also 56 Versuchsbedingungen. Die Itemanzahl jeder Versuchsbedingung und die Anzahl an getesteten Probanden pro Versuchsbedingung ist Anhang D zu entnehmen.

Ablauf der Erhebung

Die Erhebungen fanden im Dezember 2014 und im Januar 2015 statt und wurden von studentischen Hilfskräften und einer Studierenden der Pädagogischen Hochschule Heidelberg durchgeführt, deren wissenschaftliche Hausarbeit (Ellert, 2015) im Rahmen des Projekts angesiedelt war. Die Testleiter waren während der gesamten Sitzung anwesend und wurden vor der Erhebung umfassend geschult. Alle durch sie gegebenen Instruktionen mussten aus einem vorab entwickelten Manual vorgelesen werden, um eine maximale Standardisierung der Erhebung zu gewährleisten.

Der Ablauf der PC-Testung war wie folgt: Die erste Seite des Programms bestand aus einer Begrüßung und einer Anmeldemaske, in der die Schüler ihr Geschlecht, ihr Geburtsdatum und ihre Klassenstufe angeben konnten. Es folgten die Instruktionen zur Aufgabenstellung und zum Umgang mit dem Programm. Um sicherzustellen, dass die Instruktion verstanden wurde und das Programm bedient werden konnte, wurden zwei Übungsaufgaben präsentiert. Diese Aufgaben gehörten nicht zum pilotierenden Itempool. Daran schloss sich die eigentliche Testung an. Die Position der richtigen Antwort unter den vier Antwortalternativen wurde für jedes Item zufällig bestimmt, hierfür wurden mit der Software *R* (R Core Team, 2014) Zufallszahlen zwischen 1 und 4 erzeugt. Genauere Angaben zur eingesetzten Oberfläche wurden in Kapitel 5.1 gemacht. Es wurde eine statische Testversion eingesetzt und damit weder Performanzrückmeldungen noch Feedback gegeben. Jedes Item wurde nur einmal dargeboten.

7.2.3 Datenauswertung

Die computeradministrierte Erhebung verhinderte fehlende Werte, es lagen damit nur designbedingte fehlende Werte vor, da nicht alle Kinder alle Aufgaben bearbeiteten. Diese mussten nicht weiter aufbereitet werden, da sie bei der Raschskalierung explizit mitberücksichtigt wurden und bei allen Berechnungen im Sinne der klassischen Testtheorie keine Rolle spielten. Die Daten wurden auf unplausible Reaktionszeiten hin inspiziert. Hohe Reaktionszeiten sind hierbei als weniger problematisch anzusehen. Wenn

Kinder ausgeschlossen worden wären, die sehr lange für die Bearbeitung der Aufgaben benötigten, dann würden die Ergebnisse der Pilotierung und damit die Itemselektion auf Basis der Daten erfolgen, die von eher schnell arbeitenden Kindern stammen. Somit wäre die vorläufige Testendversion optimiert für eine Population mit einer eher höheren Arbeitsgeschwindigkeit. Die Zielgruppe des Tests sollte jedoch auch Kinder umfassen, die keine hohe Arbeitsgeschwindigkeit aufweisen, etwa weil sie leistungsschwach oder weniger konzentriert sind. Damit wäre ein Ausschluss der *Trials* mit auffallend langen Reaktionszeiten dem Ziel der Pilotierung nicht dienlich. Auffallend niedrige Reaktionszeiten in der Testung können jedoch als unplausibel angesehen werden, da diese Reaktionszeiten auf eine weniger ernsthafte Testbearbeitung hindeuten (Proband hat sich „durchgeklickt“). Die in Kapitel 8.1.2.3 gemachten Angaben zur Identifikation extrem kleiner Reaktionszeiten treffen analog auf die Pilotierungsstudie zu. Insgesamt wurden nicht mehr als 1 % der *Trials* entfernt. Es wurden häufig Durchgänge derselben Probanden ausgeschlossen, was für die Validität der hier durchgeführten Extremwertanalyse spricht. Diese *Trials* wurden für die weiteren Analysen nicht berücksichtigt. Sie sind jedoch nicht als fehlende Werte im eigentlichen Sinne anzusehen, vielmehr wurden sie auf Grund ihrer geringen Repräsentativität für die spätere Testsituation ausgeschlossen.

Die Berechnung der Kennwerte der klassischen Testtheorie (KTT) erfolgte anhand der in Pospeschill (2010) angeführten Formeln. Die im Rahmen der *Item-Response*-Theorie nötigen Berechnungen zur Eindimensionalität und Reliabilität wurden mit dem Programm *ConQuest* (Wu, Adams, Wilson & Haldane, 2007) ausgeführt. Reliabilitäten wurden als Cronbachs Alpha berechnet. Wurde die theoretische Annahme der Eindimensionalität in den Daten nicht erreicht, so wurden die besonders stark von der Skala abweichenden Items ausgeschlossen und für die verbleibenden Items erneut eine Prüfung auf Eindimensionalität durchgeführt. Als Richtwerte zur Bewertung auffälliger Kenngrößen dienen *T*-Werte, die kleiner −2.0 und größer 2.0 waren (Wu et al., 2007). Bei den Berechnungen im Rahmen der probabilistischen Testtheorie wurden alle Items und Personen in Parameter überführt, die auf einer gemeinsamen Logit-Skala angeordnet sind (Einhaus &

Schecker, 2007, S. 156), die entsprechende Itemschwierigkeiten auf dieser Skala wurden mit dem Kennwert δ angegeben.

7.3 Ergebnisse

Die Verteilung der Probanden auf die 56 Versuchsbedingungen und die Stichprobengröße pro Versuchsbedingung sind in Anhang D aufgeführt. Nachfolgend werden zunächst die Befunde zu den Fragestellungen P.1.1. und P.1.2. vorgestellt. Daran schließen sich die Ergebnisse zu den unter P.2. subsummierten Fragestellungen P.2.1. bis P.2.4. an, die mit der klassischen Testtheorie (KTT) beantwortet werden können.

Ergebnisse zu P.1. Dimensionalitätsprüfung

Die im Kontext der *Item-Response*-Theorie (IRT) durchgeführte Itemanalyse erfolgt für jede Skala einzeln. Items mit auffälligen Kennwerten (*T*-Wert im Betrag größer 2.0) sollen aus der Skala entfernt werden, da sie keine hinreichende Passung zum Modell aufweisen. Anschließend wird die Itemanalyse mit den noch verbliebenen Items der Skala wiederholt und gegebenenfalls weitere Items eliminiert. Dieses Vorgehen ist angezeigt bei den drei Subskalen IT, IL und PL.

Bei der Analyse der Skala IT (implizite Items mit temporaler Informationsart) haben zwei Items auffällige Kennwerte (Item ITN4 mit $T=-3.1$ und MNSQ=0.74; Item ITS9 mit $T=3.2$ und MNSQ=1.39) und wurden aus der Skala entfernt.

Bei der Analyse der Skala IL (implizite Items mit lokaler Informationsart) wird Item ILS10 mit $T=2.2$ und MNSQ=1.25 aus der Skala eliminiert. Die nach der Entfernung des Items durchgeführte Itemanalyse zeigt, dass Item ILS9 ($T=2.3$; MNSQ=1.27) keine hinreichende Passung zum Modell aufweist und daher entfernt werden muss. Eine erneute Analyse der reduzierten Skala zeigt auffallende Kennwerte der Items ILN4 ($T=-2.1$; MNSQ= 0.77), ILS2 ($T=2.6$; MNSQ=1.44), ILS3 ($T=2.6$; MNSQ=1.33), ILS4 ($T=2.1$; MNSQ=1.4) und

ILS8 (*T*=3.3; MNSQ=1.41). Insgesamt sind damit drei Eliminationsvorgänge nötig und die Skala IL reduziert sich um sieben Items.

Bei der Analyse der Skala PL (paraphrasierte Items mit lokaler Informationsart) müssen sukzessive alle Items bis auf PLN2 entfernt werden, die Skala PL kann damit nicht als eindimensional angesehen werden. Eine Übersicht über die einzelnen Eliminationsschritte und den diesbezüglichen Kennwerten der einzelnen Items kann Tabelle 17 entnommen werden.

Tabelle 17: Elimination modellunkonformer Items der Skala PL

Eliminationsschritt	**Betroffenes Item**	**MNSQ gewichtet**	***T***
1	*PLS2*	1.24	2.3
2	*PLN1*	1.53	3.9
2	*PLN3*	0.73	−2.5
2	*PLN4*	0.74	−2.6
2	*PLS1*	1.38	2.5
2	*PLS3*	0.69	−2.9
2	*PLS4*	0.73	−2.4

Nachfolgende Tabellen sollen einen Überblick über die im Zuge der IRT ermittelten Skalen geben, deren Items alle als modellkonform angenommen werden können. Hierbei wird auch der Kennwert δ angeführt, welcher das Item auf der Logit-Skala verortet.

Tabelle 18: IRT-Kennwerte der Items mit Brückeninformationen

Item	***δ***	**MNSQ gewichtet**	***T***
EBN1	−0.20	1.17	1.7
EBN2	−0.24	0.89	−1.2
EBN3	−0.24	0.89	−1.2
EBN4	−0.16	0.97	−0.3
EBS1	0.48	1.06	0.6
EBS2	0.53	1.10	0.9
EBS3	0.00	1.15	1.5
EBS4	−0.16	0.90	−1.1
PBN1	−0.39	0.98	−0.2
PBN2	−0.02	1.03	0.3
PBN3	−0.28	0.96	−0.4
PBN4	0.32	1.07	0.6
PBS1	0.19	0.90	−1.0
PBS2	−0.21	1.11	1.1
PBS3	0.41	1.03	0.3
PBS4	−0.02	1.11	1.1

Tabelle 19: IRT-Kennwerte der lokalen Items

Item	***δ***	**MNSQ gewichtet**	***T***
ELN1	−0.01	1.05	0.5
ELN2	0.09	1.06	0.6
ELN3	−0.05	1.06	0.7
ELN4	−0.05	1.05	0.6
ELS1	−0.01	1.04	0.4
ELS2	−0.05	1.03	0.3
ELS3	−0.02	1.01	0.1
ELS4	0.08	1.01	0.1
ILN1	−0.44	1.00	0.0
ILN2	1.37	1.05	0.3
ILN3	−1.29	0.94	−0.5
ILS1	0.73	1.35	1.6
ILS5	0.08	1.02	0.1
ILS6	−0.08	0.83	−1.1
ILS7	−0.37	0.85	−1.0

Tabelle 20: IRT-Kennwerte der temporalen Items

Item	*δ*	MNSQ gewichtet	*T*	Item	*δ*	MNSQ gewichtet	*T*
ETN1	−0.15	1.08	0.9	*PTS3*	0.32	0.95	−0.4
ETN2	−0.30	1.06	0.7	*PTS4*	0.63	1.07	0.5
ETN3	−0.07	1.01	0.1	*ITN1*	−0.33	0.83	−1.6
ETN4	0.28	1.03	0.3	*ITN2*	−0.46	0.85	−1.4
ETS1	0.27	1.06	0.6	*ITN3*	−0.49	0.86	−1.3
ETS2	0.10	0.99	−0.1	*ITS1*	0.36	1.05	0.4
ETS3	−0.15	1.01	0.1	*ITS2*	−0.29	1.08	0.7
ETS4	0.01	0.96	−0.3	*ITS3*	0.25	1.08	0.5
PTN1	−0.08	1.00	−0.0	*ITS4*	0.08	1.13	0.9
PTN2	−0.58	1.07	0.8	*ITS5*	0.86	0.96	−0.1
PTN3	−0.54	0.98	−0.2	*ITS6*	−0.45	0.85	−1.4
PTN4	−0.64	1.00	−0.0	*ITS7*	−0.20	0.86	−1.1
PTS1	0.36	1.01	0.1	*ITS8*	0.25	1.29	1.7
PTS2	0.52	0.99	−0.0				

Tabelle 21: IRT-Kennwerte der kausalen Items

Item	*δ*	MNSQ gewichtet	*T*	Item	*δ*	MNSQ gewichtet	*T*
EKN1	−0.03	1.12	1.3	*IKN1*	−0.55	0.99	−0.1
EKN2	−0.20	0.95	−0.6	*IKN2*	−0.62	1.01	0.2
EKN3	−0.26	0.96	−0.5	*IKN3*	−0.46	1.01	0.2
EKN4	−0.16	0.95	−0.6	*IKN4*	−0.48	1.06	0.7
EKS1	0.19	1.07	0.8	*IKS1*	−0.03	0.96	−0.3
EKS2	0.15	1.12	1.2	*IKS2*	−0.63	0.97	−0.3
EKS3	0.47	1.10	0.9	*IKS3*	−0.02	0.96	−0.3
EKS4	−0.16	0.97	−0.3	*IKS4*	0.09	0.99	−0.0
PKN1	0.17	0.98	−0.2	*IKS5*	−0.28	0.96	−0.4
PKN2	−0.34	1.10	1.1	*IKS6*	−0.09	1.00	0.0
PKN3	0.03	1.01	0.1	*IKS7*	1.25	1.01	0.1
PKN4	−0.41	1.05	0.5	*IKS8*	0.09	1.04	0.4
PKS1	0.22	1.08	0.8	*IKS9*	0.09	1.03	0.2
PKS2	0.31	0.96	−0.3	*IKS10*	1.01	0.98	−0.0
PKS3	−0.05	0.97	−0.3	*IKS11*	−0.12	1.04	0.4
PKS4	0.08	0.93	−0.7	*IKS12*	0.74	1.03	0.2

Zusammengefasst ergeben sich somit die in Tabelle 22 aufgeführten Reliabilitäten und Befunde zur Eindimensionalität (Fragestellungen P.1.1. und P.1.2.) Die Skala PL kann nicht als eindimensional angesehen werden. Die Skala IK ist zwar in sich homogen, jedoch nur von eingeschränkter Reliabilität. Die Skala PT weist Eindimensionalität auf und hat eine eher niedrige Reliabilität. Alle anderen Skalen sind eindimensional und von akzeptabler Reliabilität, ihr Cronbachs α ist größer .70. In der Skala IT mussten die Items ITN4 und ITS9 eliminiert werden, in der Skala IL die Items ILS2-4, ILS8-10 und ILN4. Der aus den einzelnen Skalen zusammengesetzte Gesamttest weist keine Eindimensionalität auf.

Tabelle 22: Ergebnisse der Dimensionalitäts- und Reliabilitätsprüfung der Skalen

Skala	**Eindimen-sionalität**	**Reliabilität**	**Anzahl Items**	**Aus Skala ausgeschlossene Items**
EB	Ja	0.79	8	-
PB	Ja	0.77	8	-
EL	Ja	0.79	8	-
PL	Nein	-	1	alle außer PLN2
IL	Ja	0.76	7	ILS2-4, ILS8-10, ILN4
ET	Ja	0.77	8	-
PT	Ja	0.68	8	-
IT	Ja	0.72	12	ITN4, ITS9
EK	Ja	0.78	8	-
PK	Ja	0.78	8	-
IK	Ja	0.53	16	-
Gesamttest	Nein	0.41	108	-

Ergebnisse zu P.2. Reliabilitätsprüfung

Die Kennwerte der klassischen Testtheorie (Fragestellungen P.2.1.-P.2.4.) können den nächsten vier Tabellen entnommen werden. Die Itemschwierigkeit ist hierbei die prozentuale Wahrscheinlichkeit, das Item richtig zu lösen. Eine Itemschwierigkeit von 100 bedeutet also, dass alle Kinder, die dieses Item bearbeiteten, es richtig lösen konnten. Dies trifft auf das Item EKN3 (siehe Tabelle 26) zu, es ist mit einer Itemschwierigkeit von 100 das einfachste Item. Bei der Betrachtung der Itemschwierigkeit wurde berücksichtigt, dass ein Item auch durch Raten richtig beantwortet werden konnte. Die Itemschwierigkeiten mit Ratekorrektur können auch negative Werte annehmen. In diesem Fall sind sie nicht zu interpretieren. Analoges gilt für die Trennschärfen, die ebenfalls negativ sein können. In einem solchen Fall kann kein sinnvoller Selektionskennwert berechnet werden (Pospeschill, 2010).

Selektionskennwerte können generell als Trennschärfen verstanden werden, die um die Itemstreuung bereinigt sind (Bühner, 2011, S. 176). Der Selektionskennwert ist direkt proportional zur Trennschärfe und indirekt proportional zur Itemstreuung. Durch diese Trennschärfenkorrektur werden insbesondere die Trennschärfen jener Items stark korrigiert, deren Schwierigkeitsindizes nahe 0 oder 1 sind (Mathieu, 2014, S. 112). Items mit höheren Selektionskennwerten sind bei der Itemselektion zu bevorzugen (Fisseni, 2004, S. 43). Wenn Items auf Grund eines niedrigen Selektionskennwertes anstatt auf Grund einer niedrigen Trennschärfe nicht für die vorläufige Testendversion ausgewählt werden, dann kann der Gefahr begegnet werden, durch die Itemselektion zu viele Items mit extremen Schwierigkeiten zu verlieren (Amelang & Zielinski, 1994, S.92). Die Verwendung des Selektionskennwerts bei der Itembewertung bietet sich insbesondere bei homogenen Niveautests an (Schmidt-Atzert & Amelang, 2012).

In Tabelle 23 bis Tabelle 26 sind die ratekorrigierten Itemschwierigkeiten, die Itemvarianzen, die Trennschärfen und die Selektionskennwerte aller Items dargestellt. Negative Trennschärfen werden nachfolgend als fehlende Werte berichtet. Bei einer Itemvarianz von 0 muss ebenfalls von einer Berechnung

des Selektionskennwerts abgesehen werden. Für die Items der nicht eindimensionalen Skala PL und für alle Items, die im Zuge der Analysen zu P.1. aus den Skalen entfernt wurden, konnten keine sinnvollen Trennschärfen und Selektionskennwerte berechnet werden.

Tabelle 23: KTT-Kennwerte der Items mit Brückeninformationen

Item	**Item-schwierigkeit ratekorrigiert**	**Itemvarianz**	**Trennschärfe**	**Selektions-kennwert**
EBN1	87.879	0.083	0.590	1.026
EBN2	85.965	0.094	0.710	1.157
EBN3	88.095	0.081	0.710	1.245
EBN4	83.333	0.109	0.670	1.013
EBS1	53.086	0.228	0.590	0.618
EBS2	47.619	0.239	0.550	0.563
EBS3	75.758	0.149	0.580	0.752
EBS4	85.455	0.097	0.710	1.139
PBN1	97.701	0.017	0.670	2.574
PBN2	76.608	0.145	0.620	0.815
PBN3	92.982	0.050	0.680	1.523
PBN4	53.333	0.228	0.560	0.587
PBS1	63.218	0.200	0.670	0.750
PBS2	78.495	0.135	0.580	0.788
PBS3	48.889	0.236	0.590	0.607
PBS4	69.399	0.177	0.570	0.678

Tabelle 24: KTT-Kennwerte der lokalen Items

Item	Item-schwierigkeit ratekorrigiert	Itemvarianz	Trennschärfe	Selektions-kennwert	Item	Item-schwierigkeit ratekorrigiert	Itemvarianz	Trennschärfe	Selektions-kennwert
ELN1	95.628	0.032	0.640	1.797	*PLS4*	53.086	0.228	-	-
ELN2	86.885	0.089	0.630	1.058	*ILN1*	27.684	0.248	0.700	0.703
ELN3	97.849	0.016	0.620	2.461	*ILN2*	−14.620	0.121	0.410	0.590
ELN4	93.651	0.045	0.630	1.479	*ILN3*	57.778	0.216	0.780	0.838
ELS1	95.628	0.032	0.640	1.797	*ILN4*	54.802	0.224	-	-
ELS2	95.699	0.031	0.650	1.839	*ILS1*	−04.444	0.170	0.320	0.388
ELS3	89.583	0.072	0.650	1.211	*ILS2*	17.778	0.236	-	-
ELS4	85.185	0.099	0.640	1.018	*ILS3*	50.282	0.234	-	-
PLN1	54.023	0.226	-	-	*ILS4*	11.864	0.224	-	-
PLN2	−15.254	0.117	-	-	*ILS5*	13.450	0.228	0.660	0.691
PLN3	57.143	0.218	-	-	*ILS6*	17.241	0.235	0.760	0.783
PLN4	73.810	0.158	-	-	*ILS7*	25.424	0.246	0.750	0.755
PLS1	29.825	0.249	-	-	*ILS8*	57.062	0.218	-	-
PLS2	95.322	0.034	-	-	*ILS9*	56.284	0.220	-	-
PLS3	59.524	0.211	-	-	*ILS10*	56.989	0.219	-	-

Tabelle 25: KTT-Kennwerte der temporalen Items

Item	Item-schwierigkeit ratekorrigiert	Itemvarianz	Trennschärfe	Selektions-kennwert	Item	Item-schwierigkeit ratekorrigiert	Itemvarianz	Trennschärfe	Selektions-kennwert
ETN1	81.287	0.121	0.610	0.878	*PTS4*	23.497	0.245	0.430	0.435
ETN2	86.441	0.091	0.630	1.042	*ITN1*	56.410	0.220	0.730	0.778
ETN3	78.571	0.135	0.630	0.858	*ITN2*	58.788	0.214	0.710	0.768
ETN4	61.212	0.206	0.590	0.650	*ITN3*	59.524	0.211	0.700	0.761
ETS1	56.322	0.220	0.590	0.629	*ITN4*	90.303	0.067	-	-
ETS2	65.517	0.192	0.640	0.731	*ITS1*	15.789	0.233	0.370	0.384
ETS3	83.333	0.109	0.630	0.952	*ITS2*	46.199	0.241	0.400	0.408
ETS4	75.758	0.149	0.650	0.843	*ITS3*	22.424	0.243	0.330	0.335
PTN1	54.098	0.226	0.580	0.610	*ITS4*	32.075	0.250	0.300	0.300
PTN2	84.946	0.100	0.560	0.885	*ITS5*	03.268	0.199	0.480	0.538
PTN3	78.495	0.135	0.630	0.856	*ITS6*	60.494	0.209	0.720	0.788
PTN4	87.302	0.086	0.610	1.039	*ITS7*	44.242	0.243	0.680	0.689
PTS1	34.463	0.250	0.510	0.510	*ITS8*	20.468	0.241	0.090	0.092
PTS2	25.683	0.247	0.530	0.534	*ITS9*	58.621	0.214	-	-
PTS3	33.333	0.250	0.590	0.590	*ITS10*	15.152	0.231	0.030	0.031

Tabelle 26: KTT-Kennwerte der kausalen Items

Item	Item-schwierigkeit ratekorrigiert	Itemvarianz	Trennschärfe	Selektions-kennwert	Item	Item-schwierigkeit ratekorrigiert	Itemvarianz	Trennschärfe	Selektions-kennwert
EKN1	92.982	0.050	0.570	1.276	*IKN1*	84.699	0.102	0.440	0.690
EKN2	95.628	0.032	0.690	1.937	*IKN2*	87.097	0.087	0.420	0.710
EKN3	100.000	0.000	0.700	-	*IKN3*	81.287	0.121	0.340	0.489
EKN4	93.443	0.047	0.700	1.619	*IKN4*	88.095	0.081	0.280	0.491
EKS1	77.011	0.143	0.560	0.741	*IKS1*	47.541	0.239	0.460	0.471
EKS2	77.401	0.141	0.550	0.733	*IKS2*	89.247	0.074	0.470	0.863
EKS3	60.920	0.207	0.530	0.582	*IKS3*	45.355	0.242	0.450	0.458
EKS4	93.443	0.047	0.690	1.595	*IKS4*	34.392	0.250	0.360	0.360
PKN1	66.667	0.188	0.650	0.751	*IKS5*	64.444	0.196	0.470	0.531
PKN2	88.095	0.081	0.610	1.070	*IKS6*	44.974	0.242	0.380	0.386
PKN3	68.485	0.180	0.640	0.753	*IKS7*	−05.263	0.166	0.210	0.258
PKN4	95.152	0.035	0.630	1.683	*IKS8*	42.529	0.245	0.230	0.232
PKS1	58.788	0.214	0.580	0.628	*IKS9*	43.860	0.244	0.280	0.284
PKS2	52.381	0.230	0.620	0.647	*IKS10*	02.381	0.196	0.320	0.361
PKS3	71.429	0.168	0.640	0.780	*IKS11*	61.212	0.206	0.290	0.319
PKS4	66.061	0.190	0.650	0.746	*IKS12*	11.905	0.224	0.160	0.169

7.4 Interpretation und Implikation für die weitere Testentwicklung

Die in Tabelle 22 dargestellten Befunde zur Eindimensionalität und Reliabilität legen nahe, dass im Gesamttest mit allen 108 Items starke Heterogenität in Hinblick auf das erfasste Merkmal vorliegt, was so erwartet und erwünscht war. Dass dieses Merkmal ein Aggregat der 11 Skalen ist, wird deutlich durch die Befunde zur Eindimensionalität und Reliabilität dieser Skalen, die wesentlich besser sind als für den Gesamttest. Dieser Befund ist besonders bemerkenswert, da Cronbachs Alpha positiv mit der Anzahl der Items der Skala korreliert ist und spricht für die inhaltliche Bedeutung der 11 Subskalen. Insgesamt ist damit die Konstruktion eindimensionaler Skalen, die voneinander distinkte Merkmale erfassen, gelungen. Lediglich die Skala PL weist keine Eindimensionalität auf und kann daher für die vorläufige Testendversion nicht berücksichtigt werden. Die Skala IK ist homogen, allerdings nur eingeschränkt reliabel. Alle weiteren Skalen haben hinreichend gute Reliabilitäten. Die Skalen mit impliziten Items umfassen mehr Items als die anderen Skalen. Dies könnte ein Grund dafür sein, warum die Skalen weniger homogen waren (IK) bzw. aus diesen Skalen mehr Items eliminiert werden mussten (IL, IK). Generell scheint jedoch ein Ziel der Testentwicklung im Wesentlichen erreicht worden zu sein: ein Design umzusetzen, welches eine Kombination von Informations- und Aufgabenart adäquat abbildet.

Zu diesen Ergebnissen muss angemerkt werden, dass die Verlinkung der einzelnen Items miteinander nicht auf einer großen Probandenanzahl beruht (vgl. Anhang D). Damit können die mittels *ConQuest* berechneten Ergebnisse in ihrer Aussagekraft potentiell beeinträchtigt sein. Aus diesem Grund wurde für die Fragestellungen P.2.1.-P.2.4. durchgehend die klassische Testtheorie (KTT) angewendet. Die inhaltlich miteinander verbundenen Fragestellungen P.1.1. und P.1.2. sind jedoch mit den Werkzeugen der KTT nicht hinreichend zu klären.

Die Befunde der KTT legen nahe, dass die Itemschwierigkeiten sehr heterogen sind, wenn auch die meisten Items als eher einfach einzustufen sind. Die hohen Lösungswahrscheinlichkeiten sind in Hinblick auf den in der Zielpopulation

vorliegenden spezifischen Förderbedarf als positiv zu bewerten. Unterschiede zwischen den Textarten kommen nicht zum Tragen. Analog zu den Ergebnissen der probabilistischen Testtheorie sind viele Items der Skala PL nicht angemessen interpretierbar.

Auf Basis der vorliegenden Befunde konnten die Items selektiert werden, die in der vorläufigen Testendversion Anwendung finden sollen. Hierbei wurde den Empfehlungen von Pospeschill (2010) gefolgt, der zur Itemselektion schreibt: „Die Selektion der Items erfolgt anhand der *simultanen Berücksichtigung* der Ergebnisse der Itemanalyse (Itemschwierigkeit, Itemvarianz und Trennschärfe bzw. Selektionskennwert) und darüber hinausgehender Überlegungen im Hinblick auf die Reliabilität [..] und Validität […] des Tests“ (Pospeschill, 2010, S. 83).

Konkret kommen folgende Prinzipien bei der Itemauswahl zum Tragen. Die Items sollen sich bezüglich ihrer Schwierigkeit hinreichend voneinander unterscheiden, damit jedes Item einen inkrementellen diagnostischen Mehrwert bringt. Es wird nach Möglichkeit immer das Item ausgewählt, welches besonders trennscharf ist. Items mit negativen Trennschärfen werden nicht in die vorläufige Testendversion aufgenommen. Des Weiteren sollen die Durchschnittsschwierigkeiten für die einzelnen Skalen (arithmetisches Mittel) nach Möglichkeit dem theoretischen Fundament der Itementwicklung Rechnung tragen (z. B. wäre es nicht zielführend, wenn eine explizite Skala erheblich schwieriger ist als eine implizite Skala). Die Schwierigkeit der Items darf außerdem in Hinblick auf die Zielpopulation mit besonderem Förderbedarf nicht zu schwierig sein, gleichzeitig muss ein gewisses Maß an Aufgabenschwierigkeit gewährleistet sein, damit das Feedback in der dynamischen Version überhaupt zum Einsatz kommt. Darüber hinaus soll keine Skala nur aus Items bestehen, die alle dieselbe Textart haben. Items, die in ihrer Schwierigkeit nach der *Item-Response*-Theorie (IRT) zu sehr vom Fähigkeitsniveau der Probanden abweichen, sollen ebenso wenig ausgewählt werden wie Items, die die Eindimensionalität ihrer Skala gefährden. Daneben werden Items mit narrativem Aufgabenstamm besonders berücksichtigt, da sie als weniger vom Vorwissen abhängig angesehen werden können als Aufgaben

mit Sachtexten (vgl. Kapitel 3.1.3.1 und Kapitel 5.1.1). Die qualitative Voruntersuchung legte nahe, dass Kinder Informationen ausblenden, wenn diese ihren kognitiven Schemata widersprechen, es kommt also nicht zwangsläufig zu einer Akkomodation (vgl. Kapitel 6). Damit ist das Vorwissen bedeutsamer für das Antwortverhalten der Kinder als die im Aufgabenstamm gegebenen Informationen. Durch Items, deren Beantwortung weniger abhängig vom jeweiligen Vorwissen ist, soll diese Problematik in ihrem Ausmaß reduziert werden.

In Hinblick auf die Motivation der Kinder soll in der vorläufigen Testendversion die (nicht ratekorrigierte) Schwierigkeit der Items ansteigend sein, beginnend mit dem einfachsten Item (Jonkisz, Moosbrugger & Brandt, 2012). Auch soll die Testung innerhalb von einer Schulstunde durchführbar sein. Auf Grundlage der Erfahrungen in der Pilotierung soll die vorläufige Testendversion damit 33 Items umfassen, wobei alle Skalen außer PL Berücksichtigung finden sollen. Jede der zehn verbleibenden Skalen soll mit mindestens drei Items abgebildet werden. Damit ist auch eine angemessene psychometrische Qualität der Messung sichergestellt. Die auf Inferenzbildung abzielenden Skalen IL, IT und IK sollen dabei besonderes Gewicht erhalten. Sie werden mit je vier Items abgebildet.

Unter Berücksichtigung aller vorliegenden Informationen und den sich daraus ableitenden Prämissen wurde die vorläufige Testendversion erstellt. Es wurden 33 Items in die vorläufige Testendversion aufgenommen. Die in den Aufgabenstämmen der narrativen Items der vorläufigen Testendversion vorkommenden 23 Namen der Protagonisten wurden dabei so angeglichen, dass 12 Namen männlich und 11 Namen weiblich sind, je zwei dieser Namen lassen Rückschlüsse auf einen Migrationshintergrund zu. Diese Quoten wurden in Anlehnung an den Mikrozensus 2013 (Statistisches Bundesamt [Destatis], 2014a) festgelegt. Es ist nicht davon auszugehen, dass diese Namensänderungen die Aufgabenstämme der Items derart verändern, dass die Ergebnisse der qualitativen Vorerprobung und der Pilotierung nicht mehr gültig sind. Vielmehr sind sie im Sinne des *construction-integration models* als Propositionen zu sehen, die für die jeweiligen Situationsmodelle nicht von

zentraler Relevanz sind. Da sie für die richtige Lösung der Aufgabe somit nicht benötigt werden, werden sie vom Leser beim zielgerichteten Bilden mentaler Repräsentationen nicht weiter berücksichtigt (*deletion*) (vgl. Kapitel 3.3).

Alle 33 Items der vorläufigen Testendversion und ihre Reihenfolge sind in Tabelle 27 aufgelistet. Damit kann die eigentliche Testentwicklung als vorerst abgeschlossen betrachtet werden.

Tabelle 27: Items und ihre Reihenfolge in der vorläufigen Testendversion

Nummer	**Item**	**Nummer**	**Item**	**Nummer**	**Item**
1	EKN2	12	ETS4	23	PBS3
2	ELN1	13	PKN1	24	IKS1
3	PKN4	14	ETS2	25	ITS7
4	EKN4	15	PBS1	26	IKS4
5	PBN3	16	ETN4	27	PTS3
6	ELS3	17	EKS3	28	ILS7
7	EBN3	18	ILN3	29	ITS3
8	EBN2	19	ITN1	30	ILS6
9	ELS4	20	PTN1	31	ITS1
10	IKN1	21	EBS1	32	IKS10
11	PTN3	22	PKS2	33	ILN2

8 Validierung des dynamischen Lesekompetenztests

Für eine sinnvolle Anwendbarkeit des entwickelten Testverfahrens in der pädagogisch-psychologischen Praxis ist die Validierung des Instruments von zentraler Bedeutung. Ziel der Validierungsstudie ist es daher, die im dynamischen Lesekompetenztest erfassten Konstrukte in Beziehung zu externen Außenkriterien zu setzen und an Hand ihrer Zusammenhänge mit diesen spezifischen Außenkriterien abschätzen zu können, ob der Test die Konstrukte erfasst, die er messen soll. Das entwickelte Instrument wird an Grundschülern und an Kindern mit spezifischem Förderbedarf validiert.

Die Validierung wird in zwei Schritten durchgeführt. Zunächst wird der Lesekompetenztest ohne dynamische Elemente validiert, um sicherzustellen, dass die Lesekompetenz valide erhoben wird (Validierung I). Die Validierung erfolgt für die Populationen der Grundschüler (Kapitel 8.1) und die Population der Kinder mit spezifischen Förderbedarf (Kapitel 8.2) getrennt. Damit ist die Verwendung der Lesekompetenz als Validitätskriterium in Hinblick auf die Validierung der dynamischen Komponente psychometrisch hinreichend legitimiert.

Die Validierung der dynamischen Erweiterung des Lesekompetenztests (Validierung II) wird in Kapitel 8.3 (Population der Grundschüler) und in Kapitel 8.4 (Population der Kinder mit spezifischen Förderbedarf) vorgestellt, hierbei werden die konvergenten und diskriminanten Validitäten der Responsivität auf die gegebenen Hilfestellungen genauer untersucht. Dabei wird bei der Ergebnisdarstellung und -interpretation nur auf die Gesichtspunkte eingegangen, welche für die Beantwortung der jeweiligen Fragestellungen notwendig sind.

Mit der Diskussion methodischer Aspekte (Kapitel 8.5.1 bis 8.5.3) und der abschließenden Bewertung der Validierungsstudien (Kapitel 8.5.4) ist die Validierung des dynamischen Lesekompetenztests vorläufig abgeschlossen.

Eine Verortung der Qualität der validierten Testversion anhand verschiedener Testgütekriterien und den sich daraus ableitbaren nächsten möglichen Projektschritten (Kapitel 8.5.5) runden das Kapitel der Validierung ab.

8.1 Validierung der Lesekompetenzkomponente des dynamischen Lesekompetenztests an Grundschülern (Validierung Ia)

8.1.1 Fragestellung

Im Zentrum des Erkenntnisgewinns der Validierungsuntersuchung I steht der empirische Zusammenhang, den die Lesekompetenzkomponente des konstruierten Tests mit externen Außenkriterien aufweist. Für die Validität der Lesekompetenz würden hierbei Korrelationen sprechen, die im Einklang mit den bisherigen Befunden und Theorien sind. Analog zu den Befunden der einschlägigen Literatur (vgl. Kapitel 3) werden hierbei lineare Zusammenhänge betrachtet. Wenn auch das hierfür erforderliche Skalenniveau (beispielsweise bei der Notenskala) nicht immer zwangsläufig als automatisch gegeben angesehen werden kann, so hat es sich aus pragmatischen Überlegungen heraus doch in Forschung und Praxis durchgesetzt, den jeweiligen Skalen ein über Ordinalskalenniveau hinausgehendes Skalenniveau zu unterstellen (vgl. Tent, 2006). Trotz dieser skalentheoretischen Schwierigkeit sollen die in der Validierungsstudie berichteten Korrelationen aus Gründen der besseren Einordnung und Vergleichbarkeit mit anderen Befunden dieser Tradition folgen.

So kann nach Kapitel 3.1.3.1 davon ausgegangen werden, dass demografische Kontrollvariablen wie das Geschlecht und die Klassenstufe sich in den Testleitungen niederschlagen werden. Schüler der vierten Jahrgangsstufe sollten bessere Leistungen erzielen als Schüler der dritten Jahrgangsstufe, Mädchen sollten bessere Leistungen erzielen als Jungen.

Neben demografischen Kontrollvariablen sind konvergente und diskriminante Validitäten von Interesse. Die konvergente Validität zielt dabei auf die Ähnlichkeit konstruktverwandter Verfahren ab: Andere Lesekompetenzmaße

sollten in besonders starkem Maße mit den Testleitungen des computeradministrierten Lesekompetenztests zusammenhängen. Als solche kann neben der Leistung in einem etablierten, auf basale Lesefähigkeiten abzielenden Lesetest auch die Beurteilung der Leseleistung durch den Lehrer gelten. Bei beiden Maßen kann davon ausgegangen werden, dass sie nicht mit der Leistung des Lesekompetenztests identisch sind, obgleich eine hinreichende Ähnlichkeit vorhanden ist. Weniger spezifisch für das Lesen als die Lehrerbeurteilung sind dagegen allgemeine Indikatoren für schulischen Erfolg, wie beispielsweise Schulnoten. So können die Schulnoten der Fächer Deutsch und Mathematik nicht als Lesemaße im eigentlichen Sinne gelten, auch wenn Lesen selbst zu den Anforderungen gehört, die im Rahmen dieser Fächer an die Schüler gestellt werden können. Während damit spezifische Lesemaße starke Korrelationen mit der Leistung im Lesetest aufweisen sollten, ist von geringen Korrelationen der Leistung im Lesetest mit den Schulnoten in den Fächern Deutsch und Mathematik auszugehen.

Wird bei der Lehrerbeurteilung die in Deutschland übliche sechsstufige Notenskala verwendet, so gilt analog zur Skala der Schulnoten, dass eine tendenziell große Kompetenz mit einer eher kleinen Note einhergeht. Damit kann erwartet werden, dass die jeweiligen Korrelationen mit der Lesekompetenz negative Vorzeichen aufweisen.

Darüber hinaus sind die allgemeinen kognitiven Fähigkeiten für die Leistung im Lesekompetenztest von besonderer Bedeutung. Dies ist theoretisch begründbar (Kapitel 3.3) und empirisch gut belegt (Kapitel 3.1.3.1). Hohe allgemeine kognitive Fähigkeiten sollten damit mit guten Leistungen im Lesekompetenztest einhergehen.

Tendenziell eher negative Auswirkungen auf die Lesekompetenz können dagegen bei Testängstlichkeit und allgemeiner Ängstlichkeit angenommen werden. Dabei ist in Anlehnung an die Befunde der Metaanalyse von Seipp (1991) davon auszugehen, dass Testängstlichkeit stärkere negative Auswirkungen auf Testleistung hat als allgemeine Ängstlichkeit.

Ausgehend von den in Kapitel 3.1.3.1 gemachten Aussagen sind damit insbesondere folgende Relationen zwischen den einzelnen Zusammenhängen zu erwarten:

VI.1. Hypothesen zu den deskriptiven Kontrollvariablen

VI.1.1. Schüler der 4. Klasse sollten im PC-Test besser abschneiden als Schüler der 3. Klasse.

VI.1.2. Mädchen sollten im PC-Test besser abschneiden als Jungen.

VI.2. Hypothesen zu den potentiellen Einflussfaktoren der Lesekompetenz

VI.2.1. Die Leseleistung im PC-Test sollte mit der basalen Lesefähigkeit positiv korrelieren.

VI.2.2. Die Leseleistung im PC-Test sollte mit der Beurteilung der Leseleistung negativ korrelieren.

VI.2.3. Die Leseleistung im PC-Test sollte mit den allgemeinen kognitiven Fähigkeiten positiv korrelieren.

VI.2.4. Die Leistung im PC-Test sollte mit Testängstlichkeit negativ korrelieren.

VI.2.5. Die Leistung im PC-Test sollte mit allgemeiner Ängstlichkeit negativ korrelieren.

VI.3. Hypothesen zu den Indikatoren der schulischen Leistung

VI.3.1. Die Leistung im PC-Test sollte nicht stark negativ mit der Deutschnote korrelieren.

VI.3.2. Die Leistung im PC-Test sollte nicht stark negativ mit der Mathematiknote korrelieren.

8.1.2 Methodik

8.1.2.1 Stichprobe

Die Stichprobe rekrutierte sich aus neun Grundschulen aus der baden-württembergischen Region Rhein-Neckar. Keine dieser Schulen war in eine vorherige Phase des Projekts involviert. Die Analysestichprobe umfasste 169 Kinder, davon waren 89 Schüler (52.7 %) aus der dritten Jahrgangsstufe und 80 Schüler (47.3 %) aus der vierten Jahrgansstufe. Die Stichprobe bestand aus 82 Mädchen (48.5 %) und 87 Jungen (51.5 %). 154 Kinder (91.1 %) sprachen zu Hause Deutsch, 26 Kinder (15.4 %) wuchsen mehrsprachig auf. Diese Kennzahlen sind als Indikator für einen möglichen Migrationshintergrund zu verstehen. Sechs Kinder verweigerten Angaben zur Muttersprache. Das durchschnittliche Alter betrug 9 Jahre und 10 Monate (SD=0;8 Jahre). Eine Schule verweigerte die Erfassung des Alters.

Die Stichprobengrößen der einzelnen Messungen können aus mehreren Gründen von der Gesamtstichprobe von N=169 Schülern abweichen (vgl. Kapitel 8.1.2.2). Es waren zum einen nicht alle Schüler bei beiden Sitzungen anwesend, zum anderen bearbeiteten nicht alle Schüler alle Aufgaben instruktionsgemäß. Nicht alle Schüler gaben freiwillig Auskunft über ihre Schulnoten. Auch war nicht jeder Lehrer dazu bereit, die Schüler bezüglich ihrer Lesefähigkeit einzuschätzen.

8.1.2.2 Design und Ablauf der Erhebung

Die Erhebungen fanden im Juni und Juli 2015 statt. Sie wurden von studentischen Hilfskräften der Pädagogischen Hochschule Heidelberg durchgeführt, die während der gesamten Sitzungen anwesend waren. Die Hilfskräfte wurden vor der Erhebung umfassend geschult. Alle durch sie gegebenen Instruktionen mussten aus einem Manual vorgelesen werden, um eine maximale Standardisierung der Erhebung zu gewährleisten. Beide Sitzungen fanden am selben Tag statt, jede Sitzung dauerte eine Schulstunde. In einer Sitzung wurde der computeradministrierte Lesekompetenztest

durchgeführt, in der anderen Sitzung ein Testheft bearbeitet, welches so konzipiert wurde, dass es kognitive Fähigkeiten, basale Lesefähigkeiten, allgemeine Ängstlichkeit und Testängstlichkeit erfasste. Die Reihenfolge der Aufgaben im Testheft war fest determiniert, zunächst wurden die basalen Lesefähigkeiten erhoben, anschließend die kognitiven Fähigkeiten und am Ende wurden allgemeine Ängstlichkeit und Prüfungsangst gemeinsam gemessen. Die Übertragung der Daten aus dem Testheft wurde zweimal durchgeführt und beide Übertragungen miteinander abgeglichen, um die durch eine mangelhafte Qualität der Datenverarbeitung bedingten Fehler minimal zu halten. Nachfolgend sollen neben der PC-Testung auch die im Testheft verwendeten Instrumente genauer beschrieben werden. Daneben wurden die Lehrer für jeden Schüler um ihre Einschätzung der Lesekompetenz des Schülers gebeten.

Instrumente und Kennwerte

PC-Testung

Das für die PC-Testung eingesetzte Material wurde ausführlich in Kapitel 5.1 besprochen. Die eingesetzte grafische Benutzerschnittstelle wurde in Kapitel 5.1.1 skizziert.

Die Position der richtigen Antwort unter den vier Antwortalternativen wurde für jedes Item zufällig bestimmt, hierfür wurden mit der Software *R* (R Core Team, 2014) Zufallszahlen zwischen 1 und 4 erzeugt. Der Ablauf der computeradministrierten Testung war wie folgt: Die erste Seite des Programms bestand aus einer Begrüßung und einer Anmeldemaske, in der die Schüler ihr Geschlecht, ihr Geburtsdatum, ihre Klassenstufe und auf freiwilliger Basis auch ihre Note in den Fächern Mathematik und Deutsch angeben konnten. Es folgten die Instruktionen zur Aufgabenstellung und zum Umgang mit dem Programm. Um sicherzustellen, dass die Instruktion verstanden wurde und das Programm bedient werden konnte, wurde eine Übungsaufgabe (ein *Dummy*item mit dem Aufgabenstamm von Item EBN1) präsentiert. Diese Aufgabe zählte noch nicht zur eigentlichen Testung, die sich an die Übungsaufgabe anschloss.

Als Indikator für die Lesekompetenz wurde die absolute Häufigkeit richtiger Antworten verwendet. Diese kann minimal 0 und maximal 33 sein. Die im Rahmen dieser Untersuchung ermittelten Kennwerte sind Tabelle 28 zu entnehmen. Die interne Konsistenz der 33 Items beträgt α=.783 (n=127) (Cronbachs α). Sie ist damit als akzeptabel anzusehen.

Tabelle 28: Kennwerte der Skala Lesekompetenz in der Validierungsuntersuchung Ia

	Lesekompetenz
Arithmetisches Mittel	23.145
Standardabweichung	5.111
Minimum	7.000
Maximum	32.000

N=159; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 33.

Allgemeine kognitive Fähigkeiten

Wie in Kapitel 3.1.3.1 dargelegt, wird die Lesekompetenz von den allgemeinen kognitiven Fähigkeiten beeinflusst. Zur Abbildung dieser kognitiven Fähigkeiten wurde ein sprachfreies Maß gewählt, das weniger als sprachabhängige Tests durch Defizite in der Lesekompetenz negativ beeinflusst werden kann (Lohaus & Vierhaus, 2015, S. 131). Aus Gründen der Durchführungsökonomie beschränkte sich die Erhebung der kognitiven Maße auf den Subtest N1 Figurenklassifikation aus dem „Kognitiven Fähigkeitstest“ (KFT) von Heller und Perleth (2000). Durch die einfachen Figuren, mit denen in diesem Subtest gearbeitet wird, kann die kognitive Fähigkeit unabhängig von der Erfahrung im Umgang mit sprachlichem und numerischem Material erfasst werden (Stern & Neubauer, 2016). Die Art der Aufgaben ist vergleichbar mit dem verbalen Testteil, jedoch sprachfrei. Sie kommt damit dem konzeptuellen Verständnis des Lesens als „Verarbeitung verbal kodierter Problemstellungen“ (Rost & Schilling, 2006, S. 452; Rost, 1987) entgegen und zielt auf jene kognitiven Fähigkeiten ab, die keine verbal-sprachlichen Fähigkeiten im eigentlichen Sinne sind. Vielmehr weist sie eine gewisse Nähe

zu dem sprachfreieren *reasoning* auf. *Reasoning* kann als für die Lesekompetenz bedeutsam angesehen werden (vgl. Kapitel 3.1.3.1).

Bei diesem Test werden pro Item drei Figuren präsentiert, die sich alle in bestimmten Merkmalseigenschaften ähneln. Eine dazu passende Figur muss aus fünf Antwortalternativen ausgewählt werden. Die Bearbeitungszeit beträgt 9 Minuten.

Der Subtest N1 besteht aus 25 Items. Da eine Normierung des Tests nur für Jahrgangsstufe 4 vorliegt, wurde auf eine weiterführende Transformation der Kennwerte verzichtet und für alle Analysen die Summe der richtig beantworteten Items verwendet. Diese kann damit zwischen 0 und 25 liegen.

Die in dieser Untersuchung empirisch ermittelten Kennwerte der Skala sind in Tabelle 29 angegeben. Die interne Konsistenz der Items beträgt α=.706 (Cronbachs α). Sie ist damit als akzeptabel anzusehen.

Tabelle 29: Kennwerte der Skala kognitive Fähigkeiten in der Validierungsuntersuchung Ia

	Kognitive Fähigkeiten
Arithmetisches Mittel	20.080
Standardabweichung	3.315
Minimum	0.000
Maximum	25.000

N=163; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 25.

<u>Basale Lesefähigkeit</u>

Als Indikator der basalen Lesefähigkeit kam das „Salzburger Lese-Screening für die Klassenstufe 1 bis 4“ (SLS) von Mayringer und Wimmer (2005) zum Einsatz. Es wurde Form A1 verwendet.

Das Salzburger-Lesescreening ist ein Test, bei dem 70 einfache Sätze (z. B. „*Tee kann man trinken.*“ oder „*Erdbeeren sind ganz blau.*“) auf ihren Wahrheitsgehalt hin als richtig oder falsch einzuschätzen sind. Das Salzburger-Lesescreening ist ein Speedtest mit einer fest vorgegebenen Bearbeitungszeit von 3 Minuten. Es sollen möglichst viele Sätze beurteilt werden. Die Summe aller richtig bearbeiteten Sätze ergibt den vom Probanden erzielten Wert der Skala basale Lesefähigkeit. Von den Testkonstrukteuren wird eine Transformation der Testwerte in klassenstufenspezifische Kennwerte vorgeschlagen, um jahrgangsstufenspezifische Unterschiede in der Testleistung zu nihilieren. Mit dieser Transformation der Rohwerte würde sich die Rangreihe der Probanden ändern, was die Vergleichbarkeit mit dem computeradministrierten Lesekompetenztest einschränkt. Der PC-Test nimmt keine Unterscheidung zwischen den Schülern der dritten und der vierten Jahrgangsstufe vor und ist daher in der Struktur seiner Rangreihen vergleichbar mit den im SLS ermittelten Rohwerten. Eine Korrelation der im PC-Test ermittelten Lesekompetenz mit jahrgangsspezifischen Normwerten wäre damit gegenüber einer Korrelation mit den im SLS erzielten Rohwerten künstlich verringert, ohne dass sich dadurch eine Aussage über die konvergente Validität des in diesem Projekt konstruierten Lesekompetenztests ableiten lassen kann. Da diese jedoch von primärem Interesse ist, soll die computeradministrierte Lesekompetenz mit einem strukturell äquivalenten Lesemaß in Bezug gesetzt werden, den im SLS erzielten Rohwerten.

Lenhard (2013) bewertet Salzburger-Lesescreening als hoch reliabel und valide, ökonomisch und robust. Lediglich die eingeschränkte Repräsentativität der Normen und der Fokus auf basale Lesefertigkeit bei Vernachlässigung des Leseverständnisses werden von Lenhard kritisch gesehen. Die eingeschränkte Repräsentativität der Normen sind für die Validierungsstudie jedoch weniger relevant, da hier die Reihenfolge der Probanden untereinander von primärem Interesse ist, weniger der Vergleich des einzelnen Probanden mit einer Norm.

Die empirischen Kennwerte der basalen Lesefähigkeit in der hier beschriebenen Untersuchung können Tabelle 30 entnommen werden. Die Reliabilität ist laut Manual gewährleistet und beträgt .90 für die dritte

Klassenstufe und .91 für die vierte Klassenstufe (Paralleltest-Reliabilität). Von einer Berechnung der Reliabilitäten in der vorliegenden Stichprobe ist durch den Speedtest-Charakter des Tests abzusehen.

Tabelle 30: Kennwerte der Skala basale Lesefähigkeit in der Validierungsuntersuchung Ia

	Basale Lesefähigkeit
Arithmetisches Mittel	41.624
Standardabweichung	10.676
Minimum	14.000
Maximum	70.000

Anmerkung. N=165; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 70.

Testängstlichkeit und allgemeine Ängstlichkeit

Wie in Kapitel 3.1.3.1 dargelegt, kann die Lesekompetenz durch Testängstlichkeit oder durch allgemein erhöhte Ängstlichkeit beeinflusst werden. Daher ist es für empirische Untersuchungen sinnvoll, diese Variablen zu berücksichtigen.

Die Konstrukte allgemeine Ängstlichkeit und Testängstlichkeit wurden mit den Subskalen manifeste Angst und Prüfungsangst des Angstfragebogens für Schüler (AFS) von Wieczerkowski et al. (1980) erfasst. Beide Skalen umfassten 15 Items. Jedes Item besteht aus einer Aussage, der zugestimmt oder nicht zugestimmt werden kann. Wird ein Item bejaht, wird es mit einem Punkt bewertet. Die Anzahl der Punkte wird für jede Skala aufsummiert. Die so ermittelten Summen haben für beide Skalen eine Spannweite von 0 bis 15.

Von einer Überführung der Rohwerte in Normwerte wurde abgesehen. Die veraltete Normierung des Angstfragebogens war dabei für diese Entscheidung von untergeordneter Bedeutung. Vielmehr waren für die im Rahmen der vorliegenden Arbeit durchgeführten Untersuchungen die Rangreihen der

Probanden von primärem Interesse, nicht die Verortung eines Probanden in einer Vergleichspopulation.

Die empirischen Kennwerte der Skalen manifeste Angst und Prüfungsangst in der hier beschriebenen Untersuchung sind in Tabelle 31 dargestellt. Die interne Konsistenz der Skala Testangst beträgt α=.838 (Cronbachs α). Die interne Konsistenz der Skala manifeste Angst beträgt α=.800 (Cronbachs α). Beide Skalen weisen damit eine gute Reliabilität auf. Sie korrelieren mit r=.631 (p<.001; n=165; zweiseitige Testung) relativ hoch miteinander, was in Kapitel 8.5 nochmals thematisiert wird.

Tabelle 31: Kennwerte der Skalen manifeste Angst und Prüfungsangst in der Validierungsuntersuchung Ia

	Manifeste Angst	**Prüfungsangst**
Arithmetisches Mittel	5.534	5.791
Standardabweichung	3.687	3.943
Minimum	0.000	0.000
Maximum	15.000	14.000

Anmerkung. N=163; Das theoretische Minimum der Skalen liegt bei 0, das theoretische Maximum liegt bei 15.

Schulnoten in Deutsch und Mathematik und Lehrerurteil Lesen

Die zuletzt erzielten Schulnoten in den Unterrichtsfächern Deutsch und Mathematik wurden im Rahmen der computeradministrierten Lesekompetenztestung per Selbstauskunft erhoben. Selbstberichtete Schulnoten können als hinreichend reliabel angesehen werden (Feng & Rost, 2015). Die Lehrereinschätzung der Lesefähigkeit bezieht sich ebenfalls auf die Notenskala, sie kann als angemessen gutes Maß der Lesekompetenz angenommen werden (Karing, Pfost & Artelt, 2013; Südkamp, Kaiser & Möller, 2012). Während die Deutschnote neben der Leseleistung auch andere Bereiche, beispielsweise auch Leistungen in Orthographie berücksichtigt, ist das Lehrerurteil als im Besonderen spezifisch für den Kompetenzbereich Lesen anzusehen. Es waren nicht alle Schüler und Lehrer gleichermaßen bereit,

Auskunft zu den erfragten Konstrukten zu geben, so dass sich der Umfang der vorhandenen Daten zwischen den einzelnen Variablen unterscheidet. Die empirischen Kennwerte sind in Tabelle 32 dargestellt.

Tabelle 32: Kennwerte der Schulnoten in Deutsch und Mathematik und des Lehrerurteils Lesen in der Validierungsuntersuchung Ia

	Deutschnote	**Mathematiknote**	**Lehrerurteil Lesen**
Arithmetisches Mittel	2.112	1.993	2.067
Standardabweichung	0.811	0.617	0.778
Minimum	1.000	1.000	1.000
Maximum	5.000	3.500	4.000
n	76	74	89

Anmerkung. Das theoretische Minimum jeder einzelnen Skala liegt jeweils bei 1, das theoretische Maximum bei 6.

8.1.2.3 Datenauswertung

Umgang mit fehlenden und unplausiblen Daten

Der Umgang mit fehlenden und unplausiblen Daten ist in allen Validierungsuntersuchungen identisch, um eine möglichst große Ähnlichkeit und damit auch eine maximale Vergleichbarkeit zwischen den Untersuchungen zu gewährleisten.

Fehlende und unplausible Werte in der PC-Testung können sowohl bei den Reaktionszeiten (RT) als auch bei den *accuracy*-Daten (ACC) auftreten und auf mehrere Ursachen zurückgeführt werden. Ein Überblick über fehlende und unplausible Werte, ihre Diagnose und den sich daraus ableitenden Handlungsempfehlungen für die in Kapitel 8 beschriebenen Studien kann Tabelle 33 entnommen werden.

Tabelle 33: Übersicht über den Umgang mit fehlenden und unplausiblen Werten

Problem	**Betroffene Werte**	**Diagnose**	**Vorgehen**	**Begründung**
fehlende RT bei Versuch 2, wenn Item im ersten Versuch gelöst	RT	Inspektion der ACC	Kodierung als missing	designbedingt fehlend, als fehlender Wert so erwartet und auch erwünscht
instruktionswidrige Bearbeitung, Proband „klickt sich durch“, bearbeitet Item nicht	ACC	Inspektion der RT, Testleiter-notiz	Kodierung als missing	Verhalten ist weder als zufällig noch als von Lesekompetenz unabhängig anzusehen
Unterbrechung der Testung	RT	Testleiter-notiz	RT als missing kodiert, ACC als valider Wert beibehalten	Unterbrechung verändert RT, nicht aber ACC
Proband wurde nicht mit der Testung fertig (*not reached*)	ACC, RT	Testleiter-notiz	RT und ACC als missing kodiert	vgl. Kapitel 8.3.2.3: Herleitung des Indikators der Feedbackresponsivität
unsystematisch fehlende Werte durch Softwareprobleme	ACC, RT	Inspektion der Daten	Kodierung als missing	sehr geringe Fallzahl (0.201%), Mehrwert von Imputationsverfahren nicht in angemessenem Kosten-Nutzen-Verhältnis

Besonderes Augenmerk muss dabei auf jene Werte gelegt werden, die fehlen, weil sie bei der Bearbeitung ausgelassen wurden und jene, die fehlen, weil das Kind durch die Zeitbegrenzung diese Aufgabe nicht mehr erreicht hat. Rohwer (2013) spricht in diesem Zusammenhang von missing type 1 (*omitted*) und

missing type 2 *(not reached*). Bei dem vorliegenden PC-Test muss eine Antwort gegeben werden, um zur nächsten Aufgabe zu gelangen, daher kann missing type 1 im PC-Test nicht auftreten. Missing type 1 im Rätselheft wurde so behandelt, wie es die Auswertungsinstruktionen der entsprechenden Testmanuale vorschreiben. Missing type 2 trat im Rätselheft nicht auf. In der PC-Testung wurde missing type 2 einheitlich als fehlende Werte behandelt. Die Begründung hierfür ist mit dem Indikator der dynamischen Komponente des dynamischen Lesekompetenztests verbunden und Kapitel 8.3.2.3 zu entnehmen.

In diesem Zusammenhang ist instruktionswidrige Bearbeitung eine bedeutende Fehlerquelle, auf die bei der Aufbereitung und Analyse gewonnener Daten angemessen reagiert werden muss. Insbesondere liegt instruktionswidrige Bearbeitung vor, wenn der Proband sich im PC-Test „durchklickt“, also das Item derart schnell bearbeitet, dass die Validität zweifelhaft erscheint. Die Entscheidung für Cut-off-Werte in den Reaktionszeiten ist dabei immer eine vom jeweiligen Datensatz abhängige Entscheidung. Besonders kurze Reaktionszeiten sind nur in Relation zum übrigen Datensatz besonders kurz. Daher müssen auch besonders lange Reaktionszeiten im Fokus der Auswertung stehen. Um extreme Reaktionszeiten möglichst valide bestimmen zu können, werden sie für das Aggregat der Studien bestimmt, die in den Kapiteln 8.1 bis 8.4 beschrieben sind. Items, bei denen die Bearbeitung unterbrochen wurde (beispielsweise durch Toilettenpausen oder externe Störungen) wurden von der Inspektion der Reaktionszeiten ausgeschlossen.

Zunächst wurden extrem lange Reaktionszeiten identifiziert. Hierbei wurden nur Grundschüler betrachtet, da bei Schülern mit besonderem Förderbedarf verlängerte Bearbeitungszeiten nicht als ein valides Ausschlusskriterium angesehen wurden. Identifiziert wurden Durchgänge, deren Reaktionszeiten mehr als drei Interquartilsabstände vom 75 %-Perzentil der Reaktionszeiten entfernt lagen. Dies betraf insbesondere die Items an Anfang der Testung, da langsam arbeitende Schüler die letzten Items der Testung auf Grund der limitierten Bearbeitungszeit häufig nicht mehr erreichten.

Unter Ausschluss der Reaktionszeiten der so identifizierten Durchgänge wurden anschließend sowohl bei den Grundschülern als auch bei Kindern mit spezifischem Förderbedarf, den Sprachheilschülern (siehe Kapitel 8.2 und 8.4; zur Wahl dieser Population vgl. Kapitel 8.2.2.1) die Bearbeitungszeiten auf unplausibel erscheinende kleine Werte inspiziert. Der Cut-off-Wert wurde hierbei beim 1 %-Perzentil festgesetzt, das entspricht etwa zwei Probanden im Datensatz. Damit wurden Durchgänge aus den weiteren Analysen ausgeschlossen, bei denen ein Proband die minimalste oder die zweitkürzeste Reaktionszeit aufwies. Daneben kam den Testleiterbeobachtungen ein inkrementeller diagnostischer Nutzen zu. Ein vom Testleiter als problematisch (z. B. unmotiviert) aufgefallener Proband wurde bei einem bestimmten *Trial* dann ausgeschlossen, wenn er unter den fünf extremsten Werten lag. Insgesamt wurden Probanden dann komplett aus den Analysen ausgeschlossen, wenn sie mindestens sieben Mal, also bei über 20 % aller Items eine der fünf kürzesten Reaktionszeiten aufwiesen. Dies betraf ausnahmslos Kinder, die bereits den Testleitern als wenig motiviert aufgefallen waren. Sie waren alle der Population der Grundschüler zuzuordnen. Dagegen führten bei Sprachheilschülern weniger motivationale Einflüsse als vielmehr spezifische Besonderheiten zu auffälligen Reaktionszeiten, die mit der Sprachheilbeschulung in Zusammenhang gebracht werden können: die Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung oder auch die autistische Störung. Diese Störungsbilder sind häufig komorbid mit Sprachstörungen (Achhammer, 2014, S. 210; Kannengieser, 2014, S. 189). Insgesamt wurden die meisten Extremwerte bei jenen Probanden ausgemacht, die bereits a priori als auffällig eingestuft wurden. Sie sind damit nicht als zufällig und von dem in der PC-Testung erhobenen Konstrukt unabhängig anzusehen.

Daneben kam es selten zu technischen Problemen, die sich in vereinzelten fehlenden Werten der vom PC-Programm erzeugten xml-Datei niederschlug. Diese fehlenden Werte waren unsystematisch und betrafen insgesamt neun Probanden. Sie wurden als *missing* kodiert.

Insgesamt wurden 4.995 % aller *Trials* als durch unsystematische technische Schwierigkeiten oder instruktionswidrige Bearbeitung problematisch identifiziert.

In Hinblick auf den Umgang mit den aus verschiedensten Gründen als *missing* kodierten Werten muss auf Grundlage der Untersuchungsziele und unter Beachtung der Besonderheiten des jeweiligen Datensatz entschieden werden, welches Verfahren des Umgangs mit diesen Werten als besonders zielführend zu erachten ist. Für die Validierungsuntersuchungen sollte beim Umgang mit fehlenden und unplausiblen Werten von Imputationsverfahren abgesehen werden, da insbesondere modellbasierte Schätzverfahren (beispielsweise das *Full-Information-Maximum-Likelihood*-Verfahren) sehr sensitiv auf Verletzungen der diesen Verfahren zu Grunde liegenden Modellannahmen reagieren (Lüdtke, Robitzsch, Trautwein & Köller, 2007). Im Fall des vorliegenden Datensatzes würde ein Verfahren wie das *Full-Information-Maximum-Likelihood*-Verfahren auf Annahmen beruhen, welche teilweise nicht als gegeben (beispielsweise die multivariate Normalverteilung) bzw. nicht als gesichert (beispielsweise *missing at random*) angesehen werden können. Auf Basis der Empfehlungen von Lüdtke et al. (2007) zum Umgang mit fehlenden Werten wurden damit die nachfolgenden Korrelationen mit listenweisen Fallausschluss berechnet, der in Bezug auf korrelations- und regressionsbasierte Analysen den klassischen Imputationsmethoden vorgezogen werden kann (z. B. Kromrey & Hines, 1994). Die fehlenden Werte betreffen insgesamt weniger als 5 % der Trials. Der Ansatz des listenweisen Fallausschlusses ist trotz der mit ihm einhergehenden Problematiken vertretbar, wie nachfolgend dargelegt werden soll.

Die mit dem listenweisen Fallausschluss verbundene Verminderung des Stichprobenumfangs ist in dieser Fragestellung von untergeordneter Bedeutung, da die Validierung nicht auf multiplen Zusammenhängen sondern auf bivariaten Korrelationen bzw. Regressionen beruht. Die Reduktion der Stichprobengröße ist damit mit der Methode des paarweisen Fallausschlusses vergleichbar. Gleichzeitig können dadurch signifikante Ergebnisse im

besonderen Maße als „abgesichert“ gelten, da es weniger wahrscheinlich ist, dass diese allein dem großen Stichprobenumfang geschuldet sind.

Der zweite Nachteil des hier angewendeten Verfahrens ist der Verlust von Informationen, die von der Analyse ausgeschlossen werden. Damit wäre der Test lediglich validiert für Probanden, die sich nicht „durchklicken“ - eine Einschränkung die in Hinblick auf eine spätere Verwendbarkeit des Tests inhaltlich vertretbar erscheint.

Durch dieses Vorgehen beruhen die Aussagen zur Validierung auf einer Population, die vollständige Antworten produziert hat. Damit ist der Test für diese Population validiert und kann später als Grundlage dienen für Einschätzungen von Testleistungen, die keine fehlenden Werte enthalten.

8.1.3 Ergebnisse

Hypothesen zu den deskriptiven Kontrollvariablen

Zunächst sollen die Befunde zu den deskriptiven Kontrollvariablen dargelegt werden. Hierbei werden Geschlecht und Klassenstufe jeweils auf die Lesekompetenz regrediert. Da die beiden unabhängigen Variablen binär sind, wurden sie in *dummy*kodierte Variablen überführt, wobei die mit „0“ kodierte Referenzkategorie bei der Variable Geschlecht die Ausprägung „weiblich“ und bei der Variable Klassenstufe die Ausprägung „dritte Klasse“ hat. Eine Übersicht über die absoluten Häufigkeiten von Geschlecht und Klassenstufe in der Stichprobe findet sich in Tabelle 34, die Ergebnisse der Regression mit einseitiger Hypothesentestung sind Tabelle 35 zu entnehmen.

Tabelle 34: Geschlecht und Jahrgangsstufe in der Analysestichprobe Ia

	Jahrgangsstufe 3	**Jahrgangsstufe 4**
weiblich	38	44
männlich	51	36

Die Vermutung, dass Schüler der vierten Klassenstufe eine signifikant höhere Lesekompetenz als Schüler der dritten Klassenstufe aufweisen (VI.1.1.), lässt sich damit bestätigen. Ebenfalls erwartungskonform sind die Ergebnisse zur Rolle des Geschlechts: Mädchen haben eine signifikant höhere Lesekompetenz als Jungen. Der einseitige Hypothesentest auf Unterschiede zwischen Mädchen und Jungen erreicht einen *p*-Wert von .043 (VI.1.2.).

Tabelle 35: Regressionsanalysen unter Berücksichtigung der deskriptiven Kontrollvariablen Geschlecht und Klassenstufe zur Vorhersage der Lesekompetenz in der Validierungsuntersuchung Ia

Prädiktor	**Beta**	***T***	**Sig.**	**N**
Geschlecht	−.137	−1.731	.043	159
Klassenstufe	.396	5.410	<.001	159

Hypothesen zur konvergenten und diskriminanten Validität

Im Folgenden sollen nun die konvergenten und diskriminanten Validitäten der Lesekompetenz betrachtet werden. Es werden hierfür Korrelationen berichtet. Diese sind zusammenfassend in Tabelle 36 dargestellt. Alle Signifikanztests sind zweiseitig.

Tabelle 36: Zusammenhänge der Lesekompetenz mit ausgewählten Variablen in der Validierungsuntersuchung Ia

Variable	Korrelation	Sig.	n
Basale Lesefähigkeit	.512	<.001	157
Allgemeine kognitive Fähigkeiten	.404	<.001	156
Lehrerurteil Lesen	−.642	<.001	86
Deutschnote	−.117	.343	68
Mathematiknote	−.138	.270	66
Testängstlichkeit	−.191	.017	155
Allgemeine Ängstlichkeit	−.082	.313	155

Es finden sich signifikant positive Zusammenhänge der Lesekompetenz mit der basalen Lesefähigkeit (VI.2.1.) und den allgemeinen kognitiven Fähigkeiten (VI.2.3.). Lesekompetenz hängt außerdem signifikant negativ mit dem Lehrerurteil (VI.2.2.) sowie der Testängstlichkeit zusammen (VI.2.4.) und weist keinen signifikanten Zusammenhang mit der Schulnote im Fach Deutsch (VI.3.1.), der Schulnote in Mathematik (VI.3.2.) und der allgemeinen Ängstlichkeit (VI.2.5.) auf. Mit Ausnahme von VI.2.5. konnten somit alle Hypothesen bestätigt werden.

Besonders hohe Lesekompetenz findet sich demnach tendenziell eher bei Schülern, die in der Testung der basalen Lesefähigkeit und der nonverbalen kognitiven Fähigkeit jeweils hohe Testwerte erzielen und denen von Seiten der Lehrer eine gute Leseleistung bescheinigt wird. Gering ausgeprägte

Lesekompetenz findet sich hingegen eher bei Kindern mit stark ausgeprägter Testängstlichkeit.

Das negative Vorzeichen der Korrelationen der Lesekompetenz mit den Schulnoten und dem Lehrerurteil ist der verwendeten Notenskala geschuldet, bei der große Kompetenz mit einer kleinen Note einhergeht. Somit erzielen Kinder mit einer eher guten Schulnote in Deutsch bzw. in Mathematik in der Tendenz eine höhere Punktzahl im Lesekompetenztest. Schüler, denen die Lehrkräfte eine gute Note im Lesen geben, erreichen ebenfalls eine tendenziell höhere Punktzahl im Lesekompetenztest. Die negative Richtung der Zusammenhänge ist damit konform zu den Hypothesen. Der Zusammenhang wird jedoch nur beim Lehrerurteil signifikant.

8.1.4 Interpretation

Die Befunde zur konvergenten und diskriminanten Validität der Lesekompetenz sprechen für eine angemessene Validität des Konstrukts, da die Fragestellungen mit Ausnahme von VI.2.5. erwartungsgemäß beantwortet werden. Die für die Fragestellung VI.2.5. relevante Korrelation der Lesekompetenz mit allgemeiner Ängstlichkeit von $r=-.082$ ist dem von Seipp (1991) in der Metaanalyse ermitteltem Zusammenhang von $r=-.163$ nicht unähnlich. Damit spricht dieser Befund der Validierungsstudie nicht generell gegen die Validität der Lesekompetenzkomponente. Auch ihr Zusammenhang mit Testangst hat eine gewissen Ähnlichkeit zu der von Seipp (1991) gefundenen Korrelation von $r=-.233$.

Die positiven Korrelationen der computeradministrierten Lesekompetenz mit der basalen Lesefähigkeit und die negative Korrelationen mit der Lehrerbeurteilung und der Testängstlichkeit sprechen ebenso für eine valide Erfassung der Lesekompetenz wie die nicht signifikanten Korrelationen mit den Schulnoten in den Fächern Deutsch und Mathematik. Der positive Zusammenhang mit den allgemeinen kognitiven Fähigkeiten kann, wie in Kapitel 3.1.3.1 dargelegt, als gesichert angenommen werden und spricht ebenfalls für die Validität der hier erhobenen Lesekompetenz.

Insgesamt kann die in diesem Test erfasste Lesekompetenz in der Population der Grundschüler damit als hinreichend legitimiert angesehen werden.

8.2 Validierung der Lesekompetenzkomponente des dynamischen Lesekompetenztests an Schülern mit spezifischem Förderbedarf (Validierung Ib)

8.2.1 Fragestellung

Auf Basis der in Kapitel 8.1.1 gemachten Ausführungen können die Fragestellungen analog zu den Fragestellungen der Validierungsuntersuchung Ia formuliert werden. Die gegenüber der vorherigen Untersuchung veränderte Population wirkt sich nicht auf die hypothetischen Zusammenhänge der jeweiligen Variablen mit der erhobenen Lesekompetenz aus, auch wenn durch den spezifischen Förderbedarf mit verringerten Performanzen im PC-Test und in den externen Lesemaßen gerechnet werden kann.

Die Fragestellungen der Validierungsuntersuchung Ib sind damit analog zur Validierungsuntersuchung Ia.

<u>VI.1. Hypothesen zu den deskriptiven Kontrollvariablen</u>

VI.1.1. Schüler der 4. Klasse sollten im PC-Test besser abschneiden als Schüler der 3. Klasse.

VI.1.2. Mädchen sollten im PC-Test besser abschneiden als Jungen.

<u>VI.2. Hypothesen zu den potentiellen Einflussfaktoren der Lesekompetenz</u>

VI.2.1. Die Leseleistung im PC-Test sollte mit der basalen Lesefähigkeit positiv korrelieren.

VI.2.2. Die Leseleistung im PC-Test sollte mit der Beurteilung der Leseleistung negativ korrelieren.

VI.2.3. Die Leseleistung im PC-Test sollte mit den allgemeinen kognitiven Fähigkeiten positiv korrelieren.

VI.2.4. Die Leistung im PC-Test sollte mit Testängstlichkeit negativ korrelieren.

VI.2.5. Die Leistung im PC-Test sollte mit allgemeiner Ängstlichkeit negativ korrelieren.

VI.3. Hypothesen zu den Indikatoren der schulischen Leistung

VI.3.1. Die Leistung im PC-Test sollte nicht stark negativ mit der Deutschnote korrelieren.

VI.3.2. Die Leistung im PC-Test sollte nicht stark negativ mit der Mathematiknote korrelieren.

8.2.2 Methodik

8.2.2.1 Stichprobe

Für die empirische Klärung der Fragestellungen der Validierungsuntersuchung Ib wurden Kinder mit spezifischem Förderbedarf im Lesen getestet. Wie in Kapitel 3.1.3.1 dargelegt, sind insbesondere Sprachheilschüler eine für diese Erhebung besonders geeignete Population. Sie wurden daher in dieser Validierungsuntersuchung untersucht.

Bei Schülern der Sprachheilschulen liegt in Hinblick auf die Lesekompetenz spezifischer Förderbedarf vor, während gleichzeitig davon ausgegangen werden kann, dass die Lernfähigkeit auf dem Niveau der Regelschulen liegt (Guthke & Wiedl, 1996, S. 205). Damit kann das Problem möglicher Deckeneffekte (vgl. Kapitel 2.6) bei Sprachheilschülern als weniger wahrscheinlich angesehen werden.

Die Analysestichprobe umfasste 16 Kinder der dritten und vierten Jahrgangsstufe einer baden-württembergischen Sprachheilschule. Entsprechend der im Förderschulbereich häufig zu beobachtenden Geschlechtsverteilungen (Statistisches Bundesamt [Destatis], 2014b) waren davon 13 Probanden (81.3 %) männlich und 3 Probanden (18.8 %) weiblich. 11 Kinder (68.8 %) besuchten die dritte Jahrgangsstufe, 4 Schüler (31.3 %) besuchten die vierte Jahrgangsstufe. 13 Kinder (81.3 %) sprachen zu Hause Deutsch, 3 Kinder (18.8 %) wuchsen mehrsprachig auf. Diese Kennzahlen sind als Indikator für einen möglichen Migrationshintergrund zu verstehen. Das durchschnittliche Alter betrug 9 Jahre und 9 Monate (SD=0;11 Jahre).

Die Stichprobengrößen der einzelnen Messungen können aus mehreren Gründen von der Gesamtstichprobe von N=16 Schülern abweichen. Es waren zum einen nicht alle Schüler bei beiden Sitzungen anwesend, zum anderen bearbeiteten nicht alle Schüler alle Aufgaben instruktionsgemäß. Außerdem waren nicht alle Schüler und Lehrer gleichermaßen bereit, Auskunft zu den

erfragten Konstrukten der Schulnoten bzw. der Lehrereinschätzung zu geben (vgl. Kapitel 8.2.2.2).

8.2.2.2 Design und Ablauf der Erhebung

Die Erhebung fand im Juli 2015 statt. Alle Probanden wurden in Einzelsitzungen getestet. Es wurde immer zuerst der computeradministrierte Lesetest durchgeführt. Der Ablauf der Sitzung war identisch zu dem im vorherigen Kapitel beschriebenen Vorgehen. Daneben wurden analog zur Validierungsuntersuchung Ia die Schulnoten in den Fächern Deutsch und Mathematik erhoben und die Lehrer für jeden Schüler um ihre Einschätzung der Lesekompetenz des Schülers gebeten.

Instrumente und Kennwerte

Neben dem computeradministrierten Lesekompetenztest kam das in Kapitel 8.1 beschriebene Testheft in dieser Erhebung ebenfalls zum Einsatz. Es erfasste neben der basalen Lesefähigkeit auch die allgemeinen kognitiven Fähigkeiten, die allgemeine Ängstlichkeit und die Testängstlichkeit. Wie in der in Kapitel 8.1 angeführten Untersuchung wurde zusätzlich zum zu validierenden Lesekompetenztest die Bewertung der Leseleistung durch den betreffenden Lehrer erhoben und in der PC-Testung die Schulnoten in den Fächern Deutsch und Mathematik erfragt. Die Aufbereitung der erhobenen Kenngrößen erfolgte analog zur Validierungsstudie Ia. Die in dieser Untersuchung empirisch ermittelten Kenngrößen sind Tabellen 37-41 zu entnehmen.

Tabelle 37: Kennwerte der Skala Lesekompetenz in der Validierungsuntersuchung Ib

	Lesekompetenz
Arithmetisches Mittel	18.429
Standardabweichung	5.185
Minimum	11.000
Maximum	29.000

Anmerkung. N=14; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 33.

Wie aus Tabelle 37 hervorgeht, war die durchschnittliche Lesekompetenz in der computeradministrierten Testung nicht identisch zu der durchschnittlichen Lesekompetenz in der Grundschulstichprobe. Diese war mit dem arithmetischen Mittel von 23.145 weniger als 1 Standardabweichung größer als der Mittelwert der Stichprobe der Sprachheilschüler. Gleichzeitig ist die Spannweite der empirisch beobachteten Werte in der Stichprobe der Sprachheilschüler verringert. Das empirische Minimum liegt bei 11, damit wird die Lesekompetenzskala nicht vollständig ausgeschöpft. Eine systematische Varianzeinschränkung kann damit nicht ausgeschlossen werden. Die interne Konsistenz der Testitems der computeradministrierten Testung beträgt α=.791 (Cronbachs α). Sie ist damit als akzeptabel anzusehen.

Nachfolgend sind die empirischen Kennwerte der mit dem Testheft erhobenen Variablen tabellarisch dargestellt.

Tabelle 38 stellt die empirisch ermittelten Kennwerte der KFT-Subskala dar. Die interne Konsistenz der Items der Skala kognitive Fähigkeiten beträgt α=.665 (Cronbachs α). Sie ist damit als eher niedrig anzusehen. Auffallend ist, dass sich hier Deckeneffekte andeuten. Mit einem durchschnittlichen Wert von 17.188 auf der KFT-Subskala (Tabelle 38) weist die Population der Kinder mit spezifischem Förderbedarf außerdem einen Mittelwert auf, der weniger als 1 Standardabweichung unterhalb des Mittelwerts der Validierungsstudie Ia liegt.

Tabelle 38: Kennwerte der Skala kognitive Fähigkeiten in der Validierungsuntersuchung Ib

	Kognitive Fähigkeiten
Arithmetisches Mittel	17.188
Standardabweichung	3.563
Minimum	10.000
Maximum	22.000

Anmerkung. N=16; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 25.

Tabelle 39 sind die empirischen Kennwerte der basalen Lesefähigkeit in dieser Population zu entnehmen. Die Skala wird nicht ausgeschöpft, der beste Proband erreichte 47 von 70 möglichen Punkten. Dies ist unter dem Gesichtspunkt des spezifischen Förderbedarfs erwartungskonform und stellt keine Gefahr für die Gültigkeit der nachfolgenden Analysen und ihrer Ergebnisse dar. Die Reliabilität ist laut Manual gewährleistet und beträgt .90 für die dritte Klassenstufe und .91 für die vierte Klassenstufe (Paralleltest-Reliabilität). Von einer Berechnung der Reliabilitäten in der vorliegenden Stichprobe ist durch den Speedtest-Charakter des Tests abzusehen.

Tabelle 39: Kennwerte der Skala basale Lesefähigkeit in der Validierungsuntersuchung Ib

	Basale Lesefähigkeit
Arithmetisches Mittel	27.188
Standardabweichung	8.627
Minimum	12.000
Maximum	47.000

Anmerkung. N=16; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 70.

In Tabelle 40 sind die empirischen Kennwerte der Ängstlichkeitsmaße dargestellt. Wie dort ersichtlich, sind die empirischen Kennwerte der Skalen manifeste Angst und Prüfungsangst relativ gering. Die Skalen werden nicht komplett ausgeschöpft. Beide Skalen korrelieren mit r=.794 (p =.001; n=16; zweiseitige Testung) stark miteinander, was in Kapitel 8.5 nochmals thematisiert wird. Die interne Konsistenz der Items der Skala Prüfungsangst ist mit Cronbachs α=.735 akzeptabel, die interne Konsistenz der Items der Skala manifeste Angst fällt dagegen mit Cronbachs α=.532 äußerst niedrig aus. Dieser Punkt soll in der Interpretation (Kapitel 8.2.4) und der allgemeinen Diskussion der Ergebnisse (Kapitel 8.5.1) nochmals aufgegriffen werden.

Tabelle 40: Kennwerte der Skalen manifeste Angst und Prüfungsangst in der Validierungsuntersuchung Ib

	Manifeste Angst	**Prüfungsangst**
Arithmetisches Mittel	4.313	4.875
Standardabweichung	2.442	3.202
Minimum	1.000	0.000
Maximum	8.000	9.000

Anmerkung. N=16; Das theoretische Minimum der Skalen liegt bei 0, das theoretische Maximum liegt bei 15.

Tabelle 41 können die empirischen Kennwerte der Deutschnote, der Mathematiknote und des Lehrerurteils über die Lesefähigkeit entnommen werden. Die Skala Deutschnote wird nicht komplett ausgeschöpft. Abweichend zu den Populationen der Grundschüler (Kapitel 8.1.2) weisen die Lehrerbeurteilung sowie die erhobenen Schulnoten in Tabelle 41 tendenziell höhere Mittelwerte auf. Auch die Streuung ist in der Tendenz relativ zu den Grundschulstichproben erhöht, insbesondere bei der Note in Mathematik und dem Lehrerurteil. Auf Grund der geringen Stichprobenumfänge in dieser Validierungsuntersuchung soll jedoch von einer statistischen Prüfung dieser Differenzen abgesehen werden.

Tabelle 41: Kennwerte der Schulnoten in Deutsch und Mathematik und des Lehrerurteils Lesen in der Validierungsuntersuchung Ib

	Deutschnote	**Mathematiknote**	**Lehrerurteil Lesen**
Arithmetisches Mittel	2.100	3.333	3.611
Standardabweichung	0.742	2.517	1.160
Minimum	1.000	1.000	1.750
Maximum	3.000	6.000	5.000
n	5	3	9

Anmerkung. Das theoretische Minimum der Skalen liegt bei 1, das theoretische Maximum liegt bei 6.

8.2.2.3 Datenauswertung

Die Datenaufbereitung und -auswertung erfolgte analog zu dem in Kapitel 8.1.2.3 beschriebenem Vorgehen. Probanden mit geringer Motivation, die sich „durchklicken“, spielten in dieser Erhebung keine Rolle. Dies kann möglicherweise dem Setting der Einzelsitzung geschuldet sein, bei dem jeder Proband einer höheren sozialen Kontrolle durch den Testleiter unterliegt. Instruktionswidrige Bearbeitung, die eine Auswertung der Testung nicht möglich machte, waren damit nicht mangelnder Motivation geschuldet. Vielmehr spielen in dieser Population spezifische Auffälligkeiten eine Rolle, die mit der Sprachheilbeschulung in Zusammenhang gebracht werden können, wie beispielsweise die Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung oder auch die autistische Störung (vgl. Kapitel 8.1.2.3). Diese Störungsbilder sind häufig komorbid mit Sprachstörungen (Achhammer, 2014, S.210; Kannengieser, 2014, S. 189).

8.2.3 Ergebnisse

Eine Übersicht über die absoluten Häufigkeiten von Geschlecht und Klassenstufe in der Stichprobe findet sich in Tabelle 42. Auf Grund der geringen Fallzahlen soll von einer regressionsstatistischen Analyse zur Klärung der Fragestellungen VI.1.1. und VI.1.2. abgesehen werden.

Tabelle 42: Geschlecht und Jahrgangsstufe in der Analysestichprobe Ib

	Jahrgangsstufe 3	**Jahrgangsstufe 4**
weiblich	2	1
männlich	9	4

Die Korrelationen der Lesekompetenzkomponente mit den im Rahmen der Fragestellungen VI.2. und VI.3. interessierenden Variablen sind in Tabelle 43 dargestellt. Alle angegeben Signifikanzen beziehen sich auf eine zweiseitige Testung.

Bei der Betrachtung der Ergebnisse muss stets bedacht werden, dass die Stichprobengrößen so gering sind, dass auch individuelle Besonderheiten der einzelnen Probanden hier starken Einfluss auf die Befunde nehmen können. Von allgemeinen Tendenzen kann damit nicht vorbehaltslos gesprochen werden.

Lesekompetenz korreliert nach Tabelle 43 signifikant positiv mit der basalen Lesefähigkeit (VI.2.1.) und negativ mit dem Lehrerurteil Lesen (VI.2.2.). Die positive Korrelation mit den allgemeinen kognitiven Fähigkeiten (VI.2.3.) lässt sich dagegen nicht zufallskritisch absichern. Keine Signifikanz erreichen außerdem die Zusammenhänge der Lesekompetenz mit Testängstlichkeit (VI.2.4.) und allgemeiner Ängstlichkeit (VI.2.5.). Ebenfalls nicht signifikant werden die negativen Zusammenhänge der Lesekompetenz mit der Schulnote im Fach Deutsch (VI.3.1.) und der Schulnote im Fach Mathematik (VI.3.2.). Jedoch können auf Grund der geringen Stichprobengröße von weniger als zehn

Personen insbesondere die Korrelationen mit dem Lehrerurteil, der Deutschnote und der Mathematiknote noch nicht als abschließend betrachtet werden.

Tabelle 43: Zusammenhänge der Lesekompetenz mit ausgewählten Variablen in der Validierungsuntersuchung Ib

Variable	**Korrelation**	**Sig.**	**n**
Basale Lesefähigkeit	.732	.003	14
Allgemeine kognitive Fähigkeiten	.444	.112	14
Lehrerurteil Lesen	−.697	.037	9
Deutschnote	−.013	.983	5
Mathematiknote	−.217	.861	3
Testängstlichkeit	.051	.862	14
Allgemeine Ängstlichkeit	.155	.596	14

Damit geht eine hohe Performanz im computeradministrierten Lesekompetenztest mit hoher Performanz im Test der basalen Lesefähigkeit und einer guten Bewertung der Leseleistung durch den Lehrer einher. Keine andere Variable hat in dieser Population einen signifikanten Zusammenhang mit dem Lesekompetenztest, obgleich beispielsweise die Höhe der Korrelation mit den allgemeinen kognitiven Fähigkeiten mit .444 moderat hoch ausfällt.

8.2.4 Interpretation

Zusammenfassend kann somit festgehalten werden, dass die Lesekompetenzkomponente auch in dieser spezifischen Stichprobe signifikant positiv mit der basalen Lesefähigkeit und signifikant negativ mit der Lehrerbeurteilung der Leseleistung auf der sechsstufigen Notenskala korreliert. Die Richtung der Zusammenhänge ist erwartungskonform und spricht für die in Validierungsuntersuchung Ia gewonnenen Erkenntnisse, ebenso die Richtung des Zusammenhangs mit den allgemeinen kognitiven Fähigkeiten.

Da die Stichprobenumfänge in dieser Untersuchung jedoch sehr gering sind, sollten diese Ergebnisse noch nicht als abschließend abgesichert angesehen werden. Die Auswirkungen des geringen Stichprobenumfangs werden in Kapitel 8.5.2 nochmal aufgenommen und diskutiert. Daneben sprechen auch die in Kapitel 8.2.2.2 dargelegten Befunde zu den einzelnen Skalen dafür, die Ergebnisse der Untersuchung kritisch zu reflektieren. So wurden die Skalen Lesekompetenz, allgemeine kognitive Fähigkeiten, Prüfungsangst und Deutschnote nicht voll ausgeschöpft. Die damit einhergehende systematische Varianzeinschränkung kann zu einer Unterschätzung der tatsächlich vorhandenen Zusammenhänge führen. Da durch eine Erhöhung des Stichprobenumfangs die Wahrscheinlichkeit steigen kann, dass eine Skala voll ausgeschöpft wird, ist das Problem der systematischen Varianzeinschränkung als nicht unabhängig vom Problem des zu geringen Stichprobenumfangs zu sehen. Darüber hinaus weist die Skala der kognitiven Fähigkeiten eine eher niedrige Reliabilität von Cronbachs α=.665 und die Skala der allgemeinen Ängstlichkeit eine äußerst niedrige Reliabilität von Cronbachs α=.532 auf, was ebenfalls teilweise stichprobenspezifischen Besonderheiten geschuldet sein kann. Diese Problembereiche sollen in Kapitel 8.5.1 nochmals thematisiert werden.

Insgesamt kann die in diesem Test erfasste Lesekompetenz in der Population der Sprachheilschüler damit als teilweise legitimiert angesehen werden. Die im Vergleich zu der an Grundschülern durchgeführten Validierungsuntersuchung

Ia abweichenden Ergebnisse können methodischen Aspekten geschuldet sein, die in Kapitel 8.5.2 nochmals thematisiert werden.

8.3 Validierung der dynamischen Komponente des dynamischen Lesekompetenztests an Grundschülern (Validierung IIa)

8.3.1 Fragestellung

Um den statischen Lesekompetenztest zu einem dynamischen Lesekompetenztest zu erweitern, muss das Ausmaß berücksichtigt werden, in welchem die Testanden auf das erhaltene Feedback ansprechen. Dieses im Folgenden als „Feedbackresponsivität“ (FR) bezeichnete Konstrukt soll in Kapitel 8.3.2.3 explizit hergeleitet werden. Dort wird dargelegt, dass die FR unkorreliert zu der Lesekompetenzkomponente des dynamischen Lesekompetenztests ist. Daraus folgt, dass die Korrelation zwischen anderen Lesemaßen und der FR nicht bedeutsam sein kann, da die FR um die Lesekompetenz bereinigt wurde. Insbesondere ist damit der Zusammenhang der FR mit der Beurteilung der Leseleistung durch den Lehrer sowie mit der Leistung in einem Test der basalen Lesefähigkeit als unerheblich anzusehen. Des Weiteren sollten die für die Lesekompetenz erwarteten negativen Korrelationen mit Testängstlichkeit und allgemeiner Ängstlichkeit nicht zwangsläufig im selben Maße gelten. Sollte sich dieses postulierte Korrelationsmuster in den Daten finden, so ist von einer gelungenen Abgrenzung zum Konstrukt der Lesekompetenz zu sprechen, die diskriminante Validität der Feedbackresponsivität kann damit diesbezüglich als gegeben angesehen werden.

Dieser Eindruck würde sich durch eine nicht signifikante Korrelation mit den deskriptiven Kontrollvariablen Klassenstufe und Geschlecht nochmal verstärken, da beide laut theoretischer Überlegungen mit der Lesekompetenz in Zusammenhang stehen (Kapitel 3.1.3.1). Daneben lässt sich der erwartete nicht signifikante Zusammenhang zum Geschlecht auch aus Kapitel 2.1 ableiten, in dem dargelegt wurde, dass es keine signifikanten geschlechtsspezifischen Auswirkungen der dynamischen Komponente im Rahmen dynamischer

Testverfahren gibt. Wenn die Intervention aber auf Mädchen und Jungen gleich wirkt, dann ist das Ausmaß, in dem Mädchen und Jungen auf das Feedback der Testung mit einer Leistungssteigerung reagieren beziehungsweise nicht reagieren, identisch und die Unterschiede in der Feedbackresponsivität zwischen den Geschlechtern nicht statistisch bedeutsam.

Dagegen kann im Sinne der konvergenten Validität angenommen werden, dass ein Zusammenhang zwischen der Feedbackresponsivität und den allgemeinen kognitiven Fähigkeiten besteht. Probanden mit hoher allgemeiner kognitiver Leistungsfähigkeit sollten eher von dem gegebenen Feedback profitieren. Diese Korrelation sollte jedoch nicht übermäßig hoch sein (vgl. Kapitel 2.7), wenngleich sie höher ausfallen muss als der Zusammenhang zwischen der Feedbackresponsivität und der (ebenfalls von kognitiven Faktoren mitbestimmten) basalen Lesefähigkeit.

Dagegen kann davon ausgegangen werden, dass sich der in Kapitel 2.7 dargestellte Befund replizieren lässt, dass schulischer Erfolg kaum im Zusammenhang mit der Performanz in einem dynamischen Test steht (Guthke, 1992). Die Korrelationen der Feedbackresponsivität mit den Schulnoten in den Fächern Deutsch und Mathematik sollten demnach nicht statistisch bedeutsam werden.

Analog zu den in Kapitel 8.1.1 aufgestellten Hypothesen sind damit insbesondere folgende Zusammenhänge der Feedbackresponsivität mit den spezifischen externen Außenkriterien zu erwarten:

<u>VII.1. Hypothesen zu den deskriptiven Kontrollvariablen</u>

VII.1.1. Schüler der 4. Klasse sollten eine Feedbackresponsivität zeigen, die vergleichbar mit der Feedbackresponsivität der Schüler der 3. Klasse ist.

VII.1.2. Die Responsivität auf das gegebene Feedback ist bei Mädchen und Jungen vergleichbar.

VII.2. Hypothese zu dem potentiellen Einflussfaktor der Feedbackresponsivität

VII.2.1. Die Responsivität auf das gegebene Feedback sollte mit den allgemeinen kognitiven Fähigkeiten moderat korrelieren.

VII.3. Hypothesen zu den weiteren Indikatoren

VII.3.1. Die Responsivität auf das gegebene Feedback sollte mit der basalen Lesefähigkeit nicht signifikant korrelieren.

VII.3.2. Die Responsivität auf das gegebene Feedback sollte mit der Beurteilung der Leseleistung durch den Lehrer nicht signifikant korrelieren.

VII.3.3. Die Responsivität auf das gegebene Feedback sollte mit der Testängstlichkeit nicht signifikant korrelieren.

VII.3.4. Die Responsivität auf das gegebene Feedback sollte mit allgemeiner Ängstlichkeit nicht signifikant korrelieren.

VII.3.5. Die Responsivität auf das gegebene Feedback sollte mit der Schulnote im Fach Deutsch nicht signifikant korrelieren.

VII.3.6. Die Responsivität auf das gegebene Feedback sollte mit der Schulnote im Fach Mathematik nicht signifikant korrelieren.

8.3.2 Methodik

8.3.2.1 Stichprobe

Die Stichprobe bestand aus einer zufälligen Auswahl von Schülern aus vier Grundschulen, die zufällig aus den Schulen ausgewählt wurden, die an der in Kapitel 8.1 beschriebenen Untersuchung teilnahmen. Durch diese zweifache Zufallsauswahl sollte gewährleistet werden, dass sich die beiden Validierungsstichproben in ihrer demografischen Beschaffenheit hinreichend ähnlich sind, Konfundierungen vorgebeugt wird und eine maximale Vergleichbarkeit zwischen beiden Validierungsuntersuchungen als gewährleistet angesehen werden kann. Dabei wurde kein Schüler zweimal getestet, er ist entweder Teil der Validierungsuntersuchung Ia oder Teil der hier beschriebenen Validierungsuntersuchung IIa.

Insgesamt wurden 59 Schüler getestet. Zwei Probanden wurden aus der Analysestichprobe ausgeschlossen, da die Responsivität auf das gegebene Feedback bei ihnen nicht bestimmt werden konnte (vgl. Kapitel 8.3.2.3). Die Analysestichprobe umfasst damit 57 Kinder. Davon waren 28 Schüler (49.1 %) aus der dritten Jahrgangsstufe und 29 Schüler (50.9 %) aus der vierten Jahrgangsstufe. Die Stichprobe bestand aus 29 Mädchen (50.9 %) und 28 Jungen (49.1 %). 57 Kinder (100 %) sprachen zu Hause Deutsch, 14 Kinder (24.6 %) wuchsen mehrsprachig auf. Diese Kennzahlen sind als Indikator für einen möglichen Migrationshintergrund zu verstehen. Das durchschnittliche Alter betrug 9 Jahre und 10 Monate (SD=0;8 Jahre).

Es waren nicht alle Schüler bei beiden Sitzungen anwesend und nicht alle Schüler bearbeiteten die Aufgaben instruktionsgemäß. Nicht alle Schüler gaben freiwillig Auskunft über ihre Schulnoten. Auch war nicht jeder Lehrer dazu bereit, die Schüler bezüglich ihrer Lesefähigkeit einzuschätzen. Die Stichprobengrößen der einzelnen Erhebungen können daher von der Analysestichprobe von N=57 Schülern abweichen (vgl. Kapitel 8.3.2.2).

8.3.2.2 Design und Ablauf der Erhebung

Der Ablauf der Sitzung war identisch zu dem in Kapitel 8.1.2 beschriebenen Vorgehen. In dieser Untersuchung kam der computeradministrierte Lesekompetenztest mit dynamischer Komponente zum Einsatz. Die Erhebungen fanden im Juni und im Juli 2015 statt.

Instrumente

Der Indikator der Feedbackresponsivität wird in Kapitel 8.3.2.3 ausführlich beschrieben. Alle weiteren verwendeten Instrumente und die sich aus diesen Instrumenten ableitenden Indizes sind analog zu Kapitel 8.1. Das dort beschriebene Testheft kam auch in dieser Untersuchung zum Einsatz, es erfasste neben der basalen Lesefähigkeit auch die allgemeinen kognitiven Fähigkeiten, die allgemeine Ängstlichkeit und die Testängstlichkeit. Entsprechend dem Vorgehen in der Validierungsstudie I wurde zusätzlich zum zu validierenden dynamischen Lesekompetenztest die Bewertung der Leseleistung durch den betreffenden Lehrer und die Schulnoten in den Fächern Deutsch und Mathematik erfragt. Die in dieser Untersuchung empirisch ermittelten Kenngrößen sind Tabellen 44-48 zu entnehmen.

Die in Tabelle 44 angeführten Kennwerte lassen den Rückschluss auf eine relativ gute Ausschöpfung der Skala Lesekompetenz zu. Die Reliabilität der Testitems des dynamischen Lesekompetenztests ist mit Cronbachs Alpha von .706 als akzeptabel zu bewerten. Die Spannweite der Skala Lesekompetenz wird relativ gut ausgeschöpft.

Tabelle 44: Kennwerte der Skala Lesekompetenz in der Validierungsuntersuchung IIa

	Lesekompetenz
Arithmetisches Mittel	22.491
Standardabweichung	5.394
Minimum	5.000
Maximum	30.000

Anmerkung. N=57; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 33.

Tabelle 45 stellt die empirischen Kennwerte der KFT-Subskala dar. Die Reliabilität der Skala ist mit Cronbachs Alpha von .708 als akzeptabel zu bewerten. Das theoretische Minimum wird in der empirischen Untersuchung nicht erreicht, der untere Wertebereich der Skala wird damit nicht ausgeschöpft.

Tabelle 45: Kennwerte der Skala kognitive Fähigkeiten in der Validierungsuntersuchung IIa

	Kognitive Fähigkeiten
Arithmetisches Mittel	20.185
Standardabweichung	2.882
Minimum	10.000
Maximum	24.000

Anmerkung. N=54; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 25.

Die empirisch ermittelten Kennwerte der basalen Lesefähigkeit sind Tabelle 46 zu entnehmen. Die Reliabilität dieser Skala ist laut Manual gewährleistet und beträgt .90 für die dritte Klassenstufe und .91 für die vierte Klassenstufe (Paralleltest-Reliabilität). Von einer Berechnung der Reliabilitäten in der vorliegenden Stichprobe ist durch den Speedtest-Charakter des Tests abzusehen.

Tabelle 46: Kennwerte der Skala basale Lesefähigkeit in der Validierungsuntersuchung IIa

	Basale Lesefähigkeit
Arithmetisches Mittel	42.825
Standardabweichung	11.477
Minimum	20.000
Maximum	69.000

Anmerkung. N=57; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 70.

Die empirischen Kenngrößen der Ängstlichkeitsmaße finden sich in Tabelle 47. Die Skala manifeste Angst weist mit α=.772 (Cronbachs Alpha) eine akzeptable Reliabilität auf, die Skala Prüfungsangst weist mit α=.837 (Cronbachs Alpha) eine gute Reliabilität auf. Beide Skalen korrelieren mit *r*=.729 (*p* <.001; n=55; zweiseitige Testung) miteinander, was in Kapitel 8.5.1 nochmals thematisiert wird.

Tabelle 47: Kennwerte der Skalen manifeste Angst und Prüfungsangst in der Validierungsuntersuchung IIa

	manifeste Angst	**Prüfungsangst**
Arithmetisches Mittel	5.327	5.750
Standardabweichung	3.445	3.905
Minimum	0.000	0.000
Maximum	12.000	14.000

Anmerkung. N=52; Das theoretische Minimum der Skalen liegt bei 0, das theoretische Maximum liegt bei 15.

In Tabelle 48 sind die empirischen Kenngrößen der Lehrerbeurteilung der Leseleistung und der Schulnoten in den Fächern Deutsch und Mathematik angeführt. Auffallend ist, dass die Notenskalen nicht ausgeschöpft werden, eine systematische Varianzeinschränkung kann damit nicht ausgeschlossen werden. Auch waren nicht alle Schüler und Lehrer gleichermaßen bereit, Auskunft zu den erfragten Konstrukten zu geben, so dass sich der Umfang der vorhandenen Daten zwischen den einzelnen Variablen unterscheidet.

Tabelle 48: Kennwerte der Schulnoten in Deutsch und Mathematik und des Lehrerurteils Lesen in der Validierungsuntersuchung IIa

	Deutschnote	**Mathematiknote**	**Lehrerurteil Lesen**
Arithmetisches Mittel	2.204	2.232	2.121
Standardabweichung	0.524	0.764	0.667
Minimum	1.500	1.000	1.000
Maximum	3.000	3.500	3.500
n	27	28	29

Anmerkung. Das theoretische Minimum der Skalen liegt bei 1, das theoretische Maximum liegt bei 6.

8.3.2.3 Datenauswertung

Umgang mit fehlenden und unplausiblen Werten

Die Behandlung fehlender und unplausibler Werte entspricht den in Kapitel 8.1.2.3 gemachten Angaben. Zusätzlich dazu wurden Reaktionszeiten im Zweitversuch nach einer falschen Antwort auf Extremwerte überprüft. Dies erfolgte analog zur Überprüfung der Reaktionszeiten im ersten Durchgang mit besonderer Berücksichtigung des verringerten Vorkommens zweiter Versuche gegenüber Erstversuchen.

Herleitung eines Indikators der Feedbackresponsivität

In Kapitel 2.5 wurden verschiedene Ansätze vorgestellt, wie die dynamische Komponente des dynamischen Tests quantifiziert werden kann. Die Ansätze von Campione und Brown beispielsweise zählen die Anzahl der benötigten Hilfestellungen, ein Ansatz, der hier nicht implementiert werden kann, da nur bei mehr als einem Feedbackdurchgang hier valide Kennwerte zu erwarten sind.

Ausgehend von den Befunden in Kapitel 8.1 und 8.2 kann angenommen werden, dass die aktuelle Lesekompetenz valide erfasst wird durch die richtigen Antworten, die ein Proband im ersten Durchgang erreicht. Damit wäre ein inverses Lesekompetenzmaß die Anzahl der falschen Antworten beim

ersten Versuch und diese ist identisch mit der Anzahl der zweiten Versuche bzw. mit der Anzahl der gegebenen Hilfestellungen. Damit kann aber die Anzahl der gegebenen Hilfestellungen nicht die Responsivität auf das Feedback abbilden, denn diese sollte sich konzeptuell von der Lesekompetenz unterscheiden.

Als theoretischer Ausgangspunkt der nachfolgenden Überlegungen diente daher der Ansatz von Budoff (vgl. Kapitel 2.5), bei dem die Performanz vor und nach der Intervention miteinander verglichen wird. Durch diesen Vergleich kann ermittelt werden, in welchem Ausmaß der Lerner von dem Training profitiert hat (*learning potential*).

Im hier zu validierenden Test folgte die Intervention immer auf eine falsche Antwort, also war die Performanz vor der Intervention/dem Feedback stets 0. Wenn im zweiten Versuch richtig geantwortet wurde, so konnte dies unter anderem an der Intervention liegen, aber man kann diese Schlussfolgerung nicht mit absoluter Sicherheit ziehen. Zufälliges Raten (im zweiten Versuch war die Ratewahrscheinlichkeit erhöht) oder reine Wiederholungseffekte könnten ebenfalls zu einer richtige Antwort im zweiten Versuch führen. Sicher ist jedoch, dass eine falsche Antwort im zweiten Versuch dann auftrat, wenn das Feedback nicht zu einer Leistungssteigerung führte. In diesem Fall ist das Ausmaß mit dem die Intervention positive Auswirkungen auf die Performanz hat, 0. Damit kann die Anzahl der falschen Antworten beim zweiten Versuch als ein negativer Indikator für die Feedbackresponsivität angesehen werden.

Gleichzeitig war die Anzahl der falschen Antworten beim zweiten Versuch aber auch davon abhängig, wie oft es überhaupt einen zweiten Versuch gab. Es lag also nahe, dass im Indikator für die Responsivität auf das Feedback die Anzahl der falschen Antworten in Relation zu der Anzahl aller zweiten Versuche gesetzt werden musste.

Dieser relative Anteil falscher Antworten unter allen Zweitversuchen war allerdings immer noch mit der Lesekompetenz konfundiert, denn auch im Zweitversuch wurde vom Testanden eine Leseleistung gefordert. Daher musste

das oben vorgestellte Maß der Lesekompetenz aus dem relativen Anteil an nicht gelösten Items auspartialisiert werden. Die Auspartialisierung wurde mittels Regressionsanalyse (Pospeschill, 2009) in *SPSS* (IBM Corporation, 2012) umgesetzt. Der relative Anteil falscher Antworten unter allen Zweitversuchen wurde regrediert auf die Lesekompetenz. Die standardisierte Differenz zwischen dem durch das Regressionsmodell vorhergesagten Wert und dem tatsächlich beobachteten Wert wurde als neue Variable abgespeichert. Diese residuale Standardisierung stellt einen Mittelwert von 0 und eine Standardabweichung von 1 sicher und erleichtert somit die Interpretation der nachfolgend durchgeführten Analysen. Sie hat keine Auswirkungen auf die inhaltliche Bedeutung des Konstrukts der Feedbackresponsivität.

Damit sei der Indikator für die Feedbackresponsivität (FR) der um die Lesekompetenz bereinigte und standardisierte relative Anteil aller falschen Antworten im zweiten Versuch.

Um die Validität dieses Index nicht zu gefährden, folgte damit für fehlende Werte, wie beispielsweise fehlende Werte des Typs *missing type 2 (not reached)* (vgl. Kapitel 8.1.2.3), dass diese nicht als falsche Antwort kodiert werden durften. Denn in diesem Fall würde ein Item mit fehlendem Wert als ein Item gewertet werden, bei dem das Feedback nicht zu einer Leistungsverbesserung führte. Tatsächlich wurde aber dem Probanden für das entsprechende Item kein Feedback gegeben, womit das Item bei diesem Probanden keinen Beitrag zum Index der Feedbackresponsivität leistete. Eine Kodierung der *accuracy*-Daten (ACC) des Items als falsche Antwort würde aber dazu führen, dass das Item ungerechtfertigterweise Einfluss auf den Index der Feedbackresponsivität nehmen würde, was die Indexvalidität in Frage stellen würde.

Die allgemeinen Kennwerte der Responsivität auf das gegebene Feedback können nachfolgender Tabelle entnommen werden. Durch die Bereinigung um die Lesekompetenz wären die nicht standardisierten Werte der Skala schwierig zu interpretieren. Durch die Standardisierung liegt der Mittelwert der Variable bei 0 und die Standardabweichung sehr nahe an 1. Für die nachfolgenden

Analysen sind die Zusammenhangsrichtungen analog zum relativen Anteil aller falschen Antworten im zweiten Durchgang zu verstehen: ein hoher Wert spricht eher dafür, dass das Feedback keinerlei Leistungsverbesserung induzierte. Ein niedriger Wert spricht dagegen für einen geringen relativen Anteil an falschen Antworten unter allen Zweitversuchen und damit für einen hohen Anteil an richtigen Antworten im zweiten Versuch.

Für zwei Probanden konnte keine Lernfähigkeit berechnet werden, da sie während der Testung keine zweiten Versuche benötigten. Sie sind in der Analysestichprobe von N=57 nicht berücksichtigt.

Tabelle 49: Kennwerte der Skala Feedbackresponsivität in der Validierungsuntersuchung IIa

	Feedbackresponsivität (FR)
Arithmetisches Mittel	0
Standardabweichung	0.991
Minimum	−1.675
Maximum	2.286

Anmerkung. N=57

8.3.3 Ergebnisse

Hypothesen zu den deskriptiven Kontrollvariablen

Analog zur Validierung der Lesekompetenz werden Geschlecht und Klassenstufe jeweils auf die Feedbackresponsivität regrediert. Da die beiden Prädiktoren binär sind, werden sie in *dummy*kodierte Variablen überführt, wobei die mit „0“ kodierte Referenzkategorie bei der Variable Geschlecht die Ausprägung „weiblich“ und bei der Variable Klassenstufe die Ausprägung „dritte Klasse“ hat. Eine Übersicht über die absoluten Häufigkeiten von Geschlecht und Klassenstufe in der Stichprobe findet sich in Tabelle 50, die Ergebnisse der Regressionen mit zweiseitigem Hypothesentest sind Tabelle 51 zu entnehmen.

Tabelle 50: Geschlecht und Jahrgangsstufe in der Analysestichprobe IIa

	Jahrgangsstufe 3	**Jahrgangsstufe 4**
weiblich	15	14
männlich	13	15

Tabelle 51: Regressionsanalysen unter Berücksichtigung der deskriptiven Kontrollvariablen Geschlecht und Klassenstufe zur Vorhersage der Feedbackresponsivität in der Validierungsuntersuchung IIa

Prädiktor	**Beta**	***T***	**Sig.**	**N**
Geschlecht	.155	1.166	.248	57
Klassenstufe	−.029	−.213	.832	57

Die Vermutung, dass Schüler der vierten Klassenstufe keine höhere Feedbackresponsivität als Schüler der dritten Klassenstufe aufweisen (VII.1.1.), lässt sich damit bestätigen. Ebenfalls erwartungskonform sind die Ergebnisse zur Rolle des Geschlechts: Mädchen und Jungen unterscheiden sich nicht signifikant in ihrer Responsivität auf das gegebene Feedback (VII.1.2.).

Hypothesen zur konvergenten und diskriminanten Validität

Im Folgenden sollen nun die konvergenten und diskriminanten Validitäten der Feedbackresponsivität (FR) betrachtet werden. Es werden dafür Korrelationen berichtet. Diese sind zusammenfassend in Tabelle 52 dargestellt. In Klammern sind die entsprechenden Kennwerte der Korrelationskoeffizienten nach Spearman angegeben, welche nicht die Linearität, sondern die Monotonie der Zusammenhänge prüfen. Damit soll versucht werden, die Art des Zusammenhangs der Feedbackresponsivität mit anderen Leistungsmaßen noch besser zu eruieren. Nachfolgende Signifikanztests sind stets zweiseitig.

Tabelle 52: Zusammenhänge der Feedbackresponsivität mit ausgewählten Variablen in der Validierungsuntersuchung IIa

Variable	**Korrelation mit FR**		**Sig.**		**n**
Lesekompetenz	.000	-	1.000	-	57
Basale Lesefähigkeit	.004	(−.120)	.977	(.374)	57
Allgemeine kognitive Fähigkeiten	−.244	(−.302)	.075	(.026)	54
Lehrerurteil Lesen	.017	(.112)	.932	(.562)	29
Deutschnote	−.003	(−.064)	.987	(.750)	27
Mathematiknote	.150	(.166)	.446	(.398)	28
Testängstlichkeit	.130	(.190)	.357	(.178)	52
Allgemeine Ängstlichkeit	−.144	(−.099)	.309	(.485)	52

Es findet sich ein negativer Zusammenhang der Feedbackresponsivität mit den durch den KFT-Subtest erfassten allgemeinen kognitiven Fähigkeiten (VII.2.1.). Probanden mit niedriger allgemeiner kognitiver Leistungsfähigkeit haben damit einen erhöhten relativen Anteil an falschen Antworten im zweiten Versuch, ihre Responsivität auf das gegebene Feedback ist damit verringert. Signifikanz erreicht lediglich die Korrelation nach Spearman, was für einen

monotonen, nicht aber für einen linearen Zusammenhang spricht. Damit ist Hypothese VII.2.1. bestätigt.

Es werden keine weiteren Korrelationen signifikant, wobei die Korrelationen nach Spearman in ihrem Muster mit den auf Linearität abzielenden Korrelationen vergleichbar sind. Damit ergibt sich folgendes Bild: Die Korrelation mit der im Test erfassten Lesekompetenz ist trivialerweise 0, was der Bereinigung der FR um diesen Faktor geschuldet ist. Sie ist hier lediglich aus Gründen der Vollständigkeit angeführt. Die basale Lesefähigkeit korreliert nicht signifikant mit der Feedbackresponsivität (VII.3.1.), ebenso die Beurteilung der Lesekompetenz durch den Lehrer (VII.3.2.). Auch die Testängstlichkeit (VII.3.3.) und die allgemeine Ängstlichkeit (VII.3.4.) weisen keine signifikanten Zusammenhänge mit der Feedbackresponsivität auf. Die Korrelation der FR mit der Deutschnote (VII.3.5.) lässt sich ebenso wenig zufallskritisch absichern wie die Korrelation mit der Mathematiknote (VII.3.6.).

8.3.4 Interpretation

Der signifikante Zusammenhang der Feedbackresponsivität mit den allgemeinen kognitiven Fähigkeiten spricht dafür, dass es sich bei der Responsivität auf das gegebene Feedback um ein Konstrukt handelt, welches zwar zu den in traditionellen Intelligenztests erfassten Konstrukten Ähnlichkeiten aufweist, jedoch nicht mit ihnen identisch ist. Dies deckt sich mit den in Kapitel 2 dargestellten Eigenschaften der dynamischen Komponente der dynamischen Testung und spricht für deren Validität. Eine inhaltliche Nähe zu den anderen Außenkriterien lässt sich empirisch nicht begründen, weder für die erhobenen schulischen Leistungen noch für die emotionalen Maße wie die Testängstlichkeit und die allgemeine Ängstlichkeit.

Obgleich alle Items der PC-Testung inhaltlich auf den Kompetenzbereich Lesen abzielen, kann das hier beschriebene Konstrukt der Feedbackresponsivität selbst nicht der Lesekompetenz zugeordnet werden. Die vorliegenden Korrelationen sprechen vielmehr dafür, dass es sich hierbei um zwei distinkte Einheiten handelt, deren Korrelationsstruktur mit denselben Außenkriterien unterschiedlich ist. Die Beobachtungen zu den deskriptiven Kontrollvariablen bestätigen dies, im Gegensatz zur Varianz in der Lesekompetenz können weder Klassenstufe noch Geschlecht signifikant Varianz in der Feedbackresponsivität vorhersagen. Damit kann das Konstrukt der Feedbackresponsivität als von der Lesekompetenz hinreichend verschieden angesehen werden.

Insgesamt kann somit die in diesem Test erfasste dynamische Komponente des dynamischen Lesekompetenztests als hinreichend abgesichert angesehen werden.

8.4 Validierung der dynamischen Komponente des dynamischen Lesekompetenztests an Schülern mit spezifischem Förderbedarf (Validierung IIb)

8.4.1 Fragestellung

Für die empirische Klärung der Fragestellungen der Validierungsuntersuchung IIb werden wie in Validierungsuntersuchung Ib Sprachheilschüler rekrutiert. Dabei wird die Feedbackresponsivität (FR) analog zur Validierungsuntersuchung IIa definiert und kann als auf dem Niveau der Regelschulen liegend angenommen werden (Guthke & Wiedl, 1996, S. 205). Die Fragestellungen der Validierungsuntersuchung IIb sind somit analog zur Validierungsuntersuchung IIa. Es ergeben sich damit folgende Fragestellungen:

<u>VII.1. Hypothesen zu den deskriptiven Kontrollvariablen</u>

VII.1.1. Schüler der 4. Klasse sollten eine Feedbackresponsivität zeigen, die vergleichbar mit der Feedbackresponsivität der Schüler der 3. Klasse ist.

VII.1.2. Die Responsivität auf das gegebene Feedback ist bei Mädchen und Jungen vergleichbar.

<u>VII.2. Hypothese zu dem potentiellen Einflussfaktor der Feedbackresponsivität</u>

VII.2.1. Die Responsivität auf das gegebene Feedback sollte mit den allgemeinen kognitiven Fähigkeiten moderat korrelieren.

<u>VII.3. Hypothesen zu den weiteren Indikatoren</u>

VII.3.1. Die Responsivität auf das gegebene Feedback sollte mit der basalen Lesefähigkeit nicht signifikant korrelieren.

VII.3.2. Die Responsivität auf das gegebene Feedback sollte mit der Beurteilung der Leseleistung durch den Lehrer nicht signifikant korrelieren.

VII.3.3. Die Responsivität auf das gegebene Feedback sollte mit der Testängstlichkeit nicht signifikant korrelieren.

VII.3.4. Die Responsivität auf das gegebene Feedback sollte mit allgemeiner Ängstlichkeit nicht signifikant korrelieren.

VII.3.5. Die Responsivität auf das gegebene Feedback sollte mit der Schulnote im Fach Deutsch nicht signifikant korrelieren.

VII.3.6. Die Responsivität auf das gegebene Feedback sollte mit der Schulnote im Fach Mathematik nicht signifikant korrelieren.

8.4.2 Methodik

8.4.2.1 Stichprobe

Insgesamt wurden 16 Kinder der dritten und vierten Jahrgangsstufe der baden-württembergischen Sprachheilschule aus Kapitel 8.2 untersucht. Eine Zuteilung der Kinder auf die Validierungsuntersuchungen Ib oder IIb erfolgte zufällig, wobei eine Ausbalancierung nach Klassenstufe und Geschlecht angestrebt wurde. Jeder Schüler wurde nur einmal getestet. Bei Schülern der Sprachheilschulen liegt in Hinblick auf die Lesekompetenz spezifischer Förderbedarf vor, während gleichzeitig davon ausgegangen werden kann, dass sich die Lernfähigkeit auf dem Niveau der Regelschulen befindet (Guthke & Wiedl, 1996, S. 205). Damit kann das Problem möglicher Deckeneffekte (vgl. Kapitel 2.6) bei Sprachheilschülern als weniger wahrscheinlich angesehen werden.

Insgesamt nahmen drei Kinder nicht an der PC-Testung teil, sie bearbeiteten nur das Testheft und finden daher keinen Eingang in die Analysestichprobe. Bei zwei Probanden konnte der PC-Test wegen instruktionswidriger Bearbeitung nicht ausgewertet werden. Bei einem Kind musste die Testung auf Grund einer vorliegenden autistischen Störung abgebrochen werden. Somit liegen nur bei 10 Probanden Daten zur Feedbackresponsivität vor, die Analysestichprobe umfasste damit 10 Kinder. Entsprechend der im Förderschulbereich häufig zu beobachtenden Geschlechtsverteilungen (Statistisches Bundesamt [Destatis], 2014b) waren davon 8 Probanden (80.0 %) männlich und 2 Probanden (20.0 %) weiblich. 6 Kinder (60.0 %) besuchten die dritte Jahrgangsstufe, 4 (40.0 %) besuchten die vierte Jahrgangsstufe. Acht Kinder (80.0 %) sprachen zu Hause Deutsch, kein Kind wuchs mehrsprachig auf. Diese Kennzahlen sind als Indikator für einen möglichen Migrationshintergrund zu verstehen. Das durchschnittliche Alter betrug 10 Jahre und 1 Monat (SD=0;8 Jahre).

Nicht alle Schüler gaben freiwillig Auskunft über ihre Schulnoten. Auch war nicht jeder Lehrer dazu bereit, die Schüler bezüglich ihrer Lesefähigkeit

einzuschätzen und für jeden Schüler eine Schullaufbahnempfehlung abzugeben. Die Stichprobengrößen der einzelnen Erhebungen können daher von der Analysestichprobe von N=10 Schülern abweichen (vgl. Kapitel 8.4.2.2).

8.4.2.2 Design und Ablauf der Erhebung

Die Erhebung fand im Juli 2015 statt. Alle Probanden wurden in Einzelsitzungen getestet. Es wurde immer zuerst der computeradministrierte Lesetest durchgeführt. Der Ablauf der Sitzung war identisch zu dem im vorherigen Kapitel beschriebenen Vorgehen. Die Zuteilung der Kinder auf die Untersuchungsstichproben Ib und IIb erfolgte zufällig, wobei eine Ausbalancierung nach Klassenstufe und Geschlecht angestrebt wurde. Daneben wurden die Schulnoten in den Fächern Deutsch und Mathematik erhoben und die Lehrer für jeden Schüler um ihre Einschätzung der Lesekompetenz des Schülers und um eine Schullaufbahnempfehlung gebeten.

Instrumente und Kennwerte

Analog zu den in Kapitel 8.3 angeführten Untersuchungen wurden zusätzlich zum zu validierenden dynamischen Lesekompetenztest die Schullaufbahnempfehlung und die Bewertung der Leseleistung durch den betreffenden Lehrer und die Schulnoten in den Fächern Deutsch und Mathematik erfragt. Das in Kapitel 8.1 beschriebene Testheft kam ebenfalls zum Einsatz, es erfasste neben der basalen Lesefähigkeit auch die allgemeinen kognitiven Fähigkeiten, die allgemeine Ängstlichkeit und die Testängstlichkeit. Die Aufbereitung der erhobenen Kenngrößen erfolgte analog zur Validierungsstudie IIa. Die in dieser Untersuchung empirisch ermittelten Kenngrößen sind Tabellen 53-59 zu entnehmen.

Die Reliabilität der Items der PC-Testung ist mit Cronbachs Alpha von .817 als gut zu bewerten. Wie aus Tabelle 53 hervorgeht, war die durchschnittliche Lesekompetenz in der computeradministrierten Testung nicht identisch zu der durchschnittlichen Lesekompetenz in der Grundschulstichprobe der Validierungsuntersuchung IIa (vgl. Kapitel 8.2.2.2). Gleichzeitig ist die

Spannweite der empirisch beobachteten Werte in der vorliegenden Stichprobe verringert. Von einer statistischen Prüfung dieser Unterschiede soll jedoch abgesehen werden, da die Stichprobengröße der hier beschriebenen Studie gering ist und somit große Konfidenzintervalle und als Folge dessen auch eine erhöhte Überlappungswahrscheinlichkeit der Konfidenzintervalle bedingt.

Die Feedbackresponsivität (Tabelle 54) weist analog zur Validierungsstichprobe II einen Mittelwert von 0 auf, was auf ihre Genese (Kapitel 8.3.2.3) zurückzuführen ist. Eine ausführliche vergleichende Betrachtung der Feedbackresponsivitäten folgt in Kapitel 8.4.3.

Tabelle 53: Kennwerte der Skala Lesekompetenz in der Validierungsuntersuchung IIb

	Lesekompetenz
Arithmetisches Mittel	14.600
Standardabweichung	4.351
Minimum	6.000
Maximum	23.000

Anmerkung. N=10; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 33.

Tabelle 54: Kennwerte der Skala Feedbackresponsivität in der Validierungsuntersuchung IIb

	Feedbackresponsivität
Arithmetisches Mittel	0.000
Standardabweichung	.943
Minimum	−1.563
Maximum	2.034

Anmerkung. N=10

Daneben wurden die Klassenleiter um eine Schullaufbahnempfehlung für jeden Schüler gebeten, der an der Testung teilnahm. Diese Erhebung erfolgte als

offene Frage, auf die auf freiwilliger Basis schriftlich geantwortet werden konnte. Es waren nicht alle Lehrkräfte bei allen Schülern gleichermaßen zur Auskunft bereit. Eine Übersicht über die genannten weiterführenden Schulen ist in Tabelle 55 ersichtlich.

Tabelle 55: Übersicht über die Schullaufbahnempfehlungen in der Validierungsuntersuchung IIb

Schullaufbahnempfehlung	**Anzahl Nennungen**
Sonderschule f. Sprachen mit Realschulzweig (Realschulabschluss)	1
Werkrealschule	3
Werkrealschule/Förderschule	1
SRH	1
Sprachheilschule*	3

Anmerkung. * Statt „Sprachheilschule“ wurde hier der konkrete Name der teilnehmenden Schule genannt. Dieser wird aus Datenschutzgründen jedoch nicht angegeben.

In Hinblick auf die Interpretation der Schullaufbahnempfehlungen kann bezüglich der Schularten generell folgendes, ansteigendes Anspruchsniveau angenommen werden: Förderschule, Hauptschule, Realschule und Gymnasium (Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany, 2015; Sliwka, 2010; Hänsel, 2003), wobei die für Baden-Württemberg spezifische Werkrealschule im an den Schüler gestellten Anspruchsniveau zwischen Haupt- und Realschule anzusiedeln ist (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2015). In den Gemeinschaftsschulen Baden-Württembergs werden jeweils Kurse auf dem Anspruchsniveau der Gymnasien, Real- und Hauptschulen angeboten (Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs, 2016).

Bei der einmal angegebenen Schulempfehlung „SRH“ handelt es sich mutmaßlich um die SRH Stephen-Hawking-Schule, eine Privatschule in der

Kinder mit und ohne körperlichen Einschränkungen verschiedene Schulabschlüsse erlangen können (SRH Stephen-Hawking-Schule Neckargemünd, 2016).

Das bereits in Kapitel 8.1.2 beschriebene Testheft kam ebenfalls zum Einsatz, es erfasste analog zu den bislang beschriebenen Validierungserhebungen neben der basalen Lesefähigkeit auch die allgemeinen kognitiven Fähigkeiten, die allgemeine Ängstlichkeit und die Testängstlichkeit. Die Aufbereitung der erhobenen Kenngrößen erfolgte wie in Kapitel 8.1.2 dargelegt. In den Tabellen 56-59 sind die empirischen Kennwerte der mit dem Testheft erhobenen Variablen tabellarisch dargestellt.

Die empirischen Kennwerte der Skala der allgemeinen kognitiven Fähigkeiten können Tabelle 56 entnommen werden. Sie legen nahe, dass eine systematische Varianzeinschränkung vorliegt. Die Skala wird nicht komplett ausgeschöpft. Mit einem durchschnittlichen Wert von 17.900 auf der KFT-Subskala weist die Population der Kinder mit spezifischem Förderbedarf einen Wert auf, der im Vergleich zur Stichprobe der Validierungsstudie IIa um weniger als 1 Standardabweichung verringert ist. Der Unterschied scheint jedoch in Hinblick auf die jeweiligen Standardabweichungen als vernachlässigbar einzustufen. Auf eine statistische Prüfung der Unterschiede wird auf Grund des geringen Stichprobenumfangs verzichtet. Die Skala weist mit α=.780 (Cronbachs Alpha) eine akzeptable Reliabilität auf.

Tabelle 56: Kennwerte der Skala kognitive Fähigkeiten in der Validierungsuntersuchung IIb

	Kognitive Fähigkeiten
Arithmetisches Mittel	17.900
Standardabweichung	3.929
Minimum	10.000
Maximum	22.000

Anmerkung. N=10; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 25.

Wie in Tabelle 57 ersichtlich, liegt die basale Lesefähigkeit der Population im Durchschnitt mehr als 1 Standardabweichung unterhalb des Mittelwerts der Stichprobe aus der entsprechenden Validierungsuntersuchung IIa (vgl. Tabelle 46). Dies ist unter dem Gesichtspunkt des spezifischen Förderbedarfs erwartungskonform und stellt keine Gefahr für die Gültigkeit der nachfolgenden Analysen und ihrer Ergebnisse dar. Auf eine zufallskritische Absicherung dieser Differenz wird auf Grund des geringen Stichprobenumfangs der Validierungsuntersuchung IIb verzichtet. Die Skala wird nicht voll ausgeschöpft, Bodeneffekte lassen sich nicht komplett ausschließen. Die basale Lesefähigkeit verfügt laut Manual über eine Paralleltest-Reliabilität von .90 für die dritte Klassenstufe und .91 für die vierte Klassenstufe. Von einer Berechnung der Reliabilitäten in der vorliegenden Stichprobe ist durch den Speedtest-Charakter des Tests abzusehen.

Tabelle 57: Kennwerte der Skala basale Lesefähigkeit in der Validierungsuntersuchung IIb

	Basale Lesefähigkeit
Arithmetisches Mittel	26.800
Standardabweichung	5.138
Minimum	21.000
Maximum	38.000

Anmerkung. N=10; Das theoretische Minimum liegt bei 0, das theoretische Maximum liegt bei 70.

In Tabelle 58 sind die empirischen Kennwerte der Ängstlichkeitsmaße dargestellt. Sie sprechen nicht für eine systematische Varianzeinschränkung. Die Reliabilitäten der Skalen sind als gut einzuschätzen, die Skala Prüfungsangst weist ein Cronbachs Alpha von α=.809 auf. Die Skala manifeste Angst weist ein Cronbachs Alpha von α=.820 auf. Beide Skalen korrelieren mit r=.834 (p=.003; n=10; zweiseitige Testung) sehr hoch miteinander, was in Kapitel 8.5.1 nochmals thematisiert wird.

Tabelle 58: Kennwerte der Skalen manifeste Angst und Prüfungsangst in der Validierungsuntersuchung IIb

	Manifeste Angst	**Prüfungsangst**
Arithmetisches Mittel	7.100	7.300
Standardabweichung	3.957	3.860
Minimum	1.000	0.000
Maximum	14.000	12.000

Anmerkung. N=10; Das theoretische Minimum der Skalen liegt bei 0, das theoretische Maximum liegt bei 15.

In Tabelle 59 finden sich die empirischen Kenngrößen der Lehrerbeurteilung der Leseleistung und der Schulnoten in den Fächern Deutsch und Mathematik. Es waren nicht alle Schüler und Lehrer gleichermaßen bereit, Auskunft zu den erfragten Konstrukten zu geben, so dass sich der Umfang der vorhandenen Daten zwischen den einzelnen Variablen unterscheidet. Abweichend zu den bisher untersuchten Populationen war es in dieser Stichprobe nicht möglich, für alle Notenskalen Standardabweichungen zu berechnen, da lediglich ein einziges Kind eine Angabe zu seiner Deutschnote machte. Die Skalen der Lehrerbeurteilung und der Mathematiknote weisen eine systematische Varianzeinschränkung auf.

Tabelle 59: Kennwerte der Schulnoten in Deutsch und Mathematik und des Lehrerurteils Lesen in der Validierungsuntersuchung IIb

	Deutschnote	**Mathematiknote**	**Lehrerurteil Lesen**
Arithmetisches Mittel	6.000	1.500	3.472
Standardabweichung	-	0.577	0.712
Minimum	6.000	1.000	2.500
Maximum	6.000	2.000	4.500
n	1	4	9

Anmerkung. Das theoretische Minimum der Skalen liegt bei 1, das theoretische Maximum liegt bei 6.

8.4.2.3 Datenauswertung

Die Datenaufbereitung und -auswertung sowie die Berechnung des Indikators der Feedbackresponsivität erfolgten analog zu dem in Kapitel 8.3.2.3 beschriebenen Vorgehen.

Probanden mit geringer Motivation, die sich „durchklicken“, spielten in dieser Erhebung keine Rolle. Dies kann möglicherweise dem Setting der Einzelsitzung geschuldet sein, bei dem jeder Proband einer höheren sozialen Kontrolle durch den Testleiter unterlag. Instruktionswidrige Bearbeitung, die eine Auswertung der Testung nicht möglich machte, waren stets den spezifischen Einschränkungen der einzelnen Kinder geschuldet, welche in einem Spannungsverhältnis zu den Anforderungen der PC-Testung standen: die Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung oder auch die autistische Störung. Diese Störungsbilder sind häufig komorbid mit Sprachstörungen (Achhammer, 2014, S.210; Kannengieser, 2014, S. 189).

8.4.3 Ergebnisse

Eine Übersicht über die absoluten Häufigkeiten von Geschlecht und Klassenstufe in der Stichprobe findet sich in Tabelle 60. Auf Grund des geringen Stichprobenumfangs kann keine statistische Prüfung der Hypothesen VII.1.1. und VII.1.2. vorgenommen werden.

Tabelle 60: Geschlecht und Jahrgangsstufe in der Analysestichprobe IIb

	Jahrgangsstufe 3	**Jahrgangsstufe 4**
weiblich	1	1
männlich	5	3

Die Korrelationen der Feedbackresponsivität sind in Tabelle 61 dargestellt. Analog zu Kapitel 8.3.3 sind die monotonen Zusammenhänge nach Spearman in Klammern angegeben. Sie unterscheiden sich nicht wesentlich von den auf Linearität abzielenden Korrelationen. Alle angegeben Signifikanzen beziehen sich auf eine zweiseitige Testung. Da die Deutschnote nur für einen einzigen Probanden vorlag, konnte ihr Zusammenhang mit der Feedbackresponsivität nicht bestimmt werden.

Auf Grund der geringen Stichprobengrößen können die Ergebnisse der Signifikanztests an dieser Stelle nicht vorbehaltslos übernommen werden, eine explorative Bewertung der jeweiligen Korrelationen bezüglich ihrer Höhe und Richtung soll daher weitere Hinweise auf die Beantwortung der Forschungsfragen geben. Bei der Betrachtung der Ergebnisse muss stets bedacht werden, dass die Stichprobengrößen so gering sind, dass auch individuelle Besonderheiten der einzelnen Probanden hier starken Einfluss auf die Befunde nehmen können. Von allgemeinen Tendenzen kann damit nicht ohne Vorbehalt gesprochen werden.

Tabelle 61: Zusammenhänge der Feedbackresponsivität mit ausgewählten Variablen in der Validierungsuntersuchung IIb

Variable	**Korrelation mit FR**		**Sig.**		**n**
Lesekompetenz	.000	-	1.000	-	10
Basale Lesefähigkeit	−.027	(.152)	.940	(.676)	10
Allgemeine kognitive Fähigkeiten	.069	(.106)	.849	(.770)	10
Lehrerurteil Lesen	.279	(.213)	.467	(.582)	9
Deutschnote	-	-	-	-	1
Mathematiknote	.678	(.447)	.322	(.553)	4
Testängstlichkeit	.278	(.275)	.437	(.441)	10
Allgemeine Ängstlichkeit	.261	(.299)	.466	(.402)	10

Wie Tabelle 61 zu entnehmen, werden keine der hier berichteten Korrelationen signifikant. Die Korrelation der Feedbackresponsivität mit der Lesekompetenz im PC-Test ist trivialerweise 0 und wird nur aus Gründen der Vollständigkeit berichtet. Der lineare Zusammenhang zwischen der Feedbackresponsivität und der basalen Lesefähigkeit ist nahe 0, der monotone Zusammenhang ist eher schwach ausgeprägt (VII.3.1.), damit lässt sich die Feedbackresponsivität hinreichend stark von der Lesekompetenzkomponente der dynamischen Testung abgrenzen.

Die Hypothese, wonach die allgemeinen kognitiven Fähigkeiten moderat mit der Feedbackresponsivität korrelieren, lässt sich weder mit Blick auf die Höhe der Korrelationen noch mit Blick auf die Ergebnisse der Signifikanztests bestätigen (VII.2.1.). Die Korrelation nach Spearman kann ebenso wenig wie die auf Linearität abzielende Korrelation zufallskritisch abgesichert werden.

Da die Monotonie des Zusammenhangs der Feedbackresponsivität mit der basalen Lesefähigkeit und mit dem Subtest des KFTs in ihrer Stärke die

Linearität übersteigt, scheint hier ein nicht-linearer Zusammenhang plausibel. Ob dieser Befund der spezifischen Verteilung der jeweiligen Skalenwerte oder einem anderen Umstand geschuldet ist, lässt sich nicht mit Sicherheit feststellen.

Die Art des Zusammenhangs zwischen der Feedbackresponsivität und der Deutschnote (VII.3.5.) kann in dieser Untersuchung nicht geklärt werden. Die Korrelationen der übrigen Notenskalen mit der Feedbackresponsivität sind größer 0. Probanden mit höheren Werten und damit schlechteren Noten auf den Skalen Lehrerurteil Lesen und Mathematiknote haben demnach einen erhöhten relativen Anteil an falschen Antworten im zweiten Versuch, ihre Responsivität auf das gegebene Feedback ist damit verringert (VII.3.2. und VII.3.6.). Die Zusammenhänge lassen sich jedoch nicht zufallskritisch absichern und sind auf Grund der sehr kleinen Stichprobengrößen von weniger als 10 Probanden mit Vorsicht zu interpretieren.

Die Korrelationen der Ängstlichkeitsmaße mit der Feedbackresponsivität zeigen ein ähnliches Bild. Eine verminderte Responsivität auf das gegebene Feedback und damit ein erhöhter Anteil an falschen Antworten im Zweitversuch findet sich andeutungsweise häufiger bei Probanden mit erhöhter Testängstlichkeit (VII.3.3.) und erhöhter allgemeiner Ängstlichkeit (VII.3.4.). Damit profitieren ängstlichere Kinder in dieser Population tendenziell weniger von den gegebenen Hilfestellungen, jedoch werden diese Zusammenhänge nicht signifikant. Sie sind noch nicht als abschließend abgesichert anzusehen.

Zum Zusammenhang zwischen der Feedbackresponsivität und den allgemeinen kognitiven Fähigkeiten

Da die Befunde von eingeschränkter Aussagekraft sind, soll nachfolgend versucht werden, weitere Informationen über die Validität der Feedbackresponsivität zu erhalten, um über die Hypothesen hinaus mehr Erkenntnisse über dieses noch wenig beforschte Konstrukt zu erhalten. In diesem Zusammenhang ist die postulierte Korrelation zwischen den allgemeinen kognitiven Fähigkeiten und der Feedbackresponsivität von besonderer Bedeutung, da sich hier Differenzen zwischen den Befunden der

Validierungsuntersuchungen IIa und IIb in besonderem Maße zeigen. Daher sollen im nächsten Schritt Grund- und Sprachheilschüler in Hinblick auf ihre allgemeinen kognitiven Fähigkeiten und ihre Feedbackresponsivität miteinander in Bezug gesetzt werden. Falls es hier Auffälligkeiten gibt, so könnten sie die verringerte Korrelation der Feedbackresponsivität in der Population der Sprachheilschüler möglicherweise teilweise erklären. Dabei ist es zu kurz gegriffen, nur Mittelwerte auf Unterschiede zu prüfen. Vielmehr gilt es, die Struktur der Daten im Ganzen miteinander zu vergleichen, um durch eine umfassende Exploration zu einer genaueren Schlussfolgerung zu kommen. Hierbei müssen auch Verteilungsmerkmale wie die Spannweiten, die Schiefen oder die Dichten in den beiden Gruppen mitberücksichtigt werden. So bietet sich beispielsweise eine grafische Exploration mittels *Boxplot* (Fahrmeir, Künstler, Pigeot & Tutz, 2011) an. Damit kann neben der Streuung auch die Dispersion der Daten berücksichtigt werden. Abbildung 12 zeigt die Verteilung der Feedbackresponsivität (FR) in den beiden Populationen. Die FR muss für einen sinnvollen Vergleich hierbei für die gepoolte Stichprobe aus Grundschülern und Sprachheilschülern berechnet werden.

Der Abbildung kann entnommen werden, dass sich die Interquartilsabstände beider Populationen relativ ähnlich sind, beide bewegen sich zwischen −1 und +1 und überlappen sich in ihrem Wertebereich. Auch die Mediane sind vergleichbar. Die Stichprobe der Sprachheilschüler hat eine kleinere Spannweite und scheint weniger symmetrisch in der Verteilung der FR-Werte zu sein, was zu einem gewissen Teil sicher auch ihrem kleineren Stichprobenumfang geschuldet ist. Die Dichte der Verteilung ist im Bereich zwischen unterem Quartil und Median erhöht, wohingegen in der Population der Grundschüler die Dichte im Bereich zwischen Median und oberem Quartil höher ist. Dennoch scheinen sich beide Populationen in ihren Verteilungen nicht auffällig voneinander zu unterscheiden. Dies entspricht der Erwartung, dass die in dieser Untersuchung ermittelte Feedbackresponsivität vergleichbar sein sollte mit der Feedbackresponsivität, die in der Validierungsuntersuchung IIa an Grundschülern erhoben sein sollte (vgl. Guthke & Wiedl, 1996, S. 205).

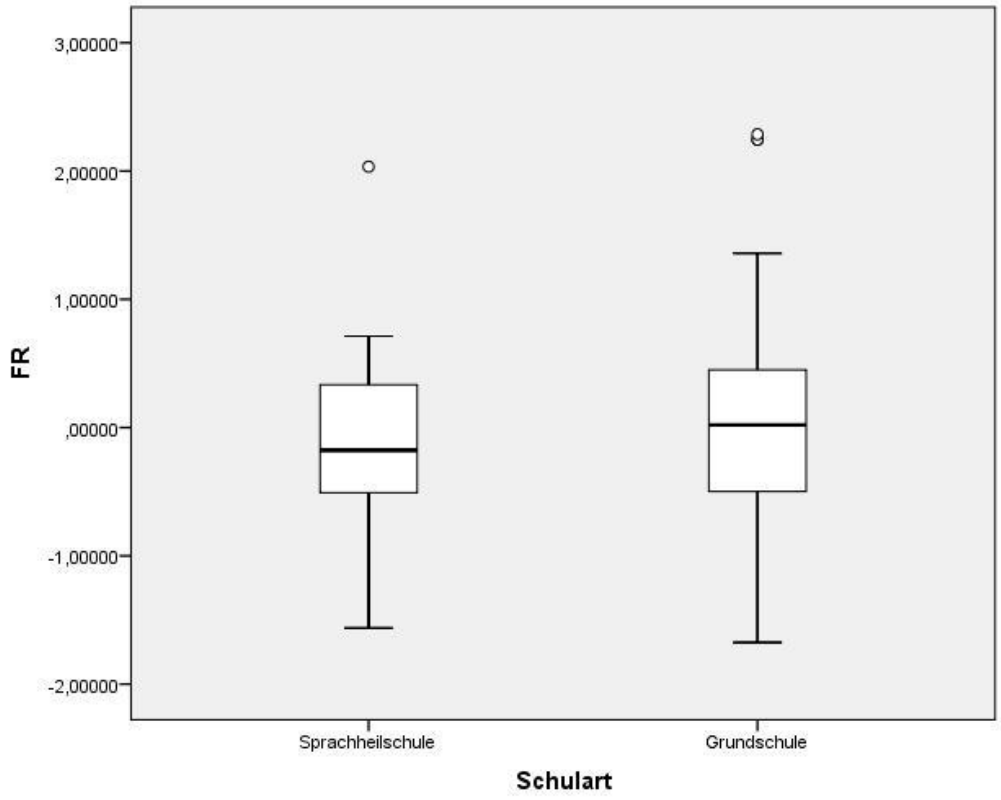

Abbildung 12: Validierung: Feedbackresponsivität (FR) in Abhängigkeit von der Schulart

Die allgemeinen kognitiven Fähigkeiten der beiden Populationen weisen dagegen ein davon abweichendes Verteilungsmuster auf, wie in Abbildung 13 deutlich wird. Während sich die Mediane wenig unterscheiden, so liegt doch das obere Quartil der Kinder der Sprachheilschule unter dem Median der Grundschüler. Gleichzeitig finden sich aber auch vereinzelt bei den Grundschülern stark nach unten abweichende Werte in der KFT-Subskala. Diese extremen Werte verringern das arithmetische Mittel der Grundschulpopulation, so dass der reine Mittelwertsunterschied (vgl. Kapitel 8.3.2.2 und Kapitel 8.4.2.2) damit weniger aussagekräftig wird. Für einen Unterschied in den Verteilungsstrukturen spricht darüber hinaus auch die in der Population der Sprachheilschüler ähnlich große Spannweite der empirischen Werte bei einer gleichzeitig weniger stark ausgeprägter Symmetrie der Verteilung. Diese kann nicht auf eine große Anzahl an Probanden mit Sprachheilbeschulung zurückgeführt werden. Damit könnte das Ausbleiben des erwarteten Zusammenhangs zwischen der FR und den allgemeinen kognitiven

Fähigkeiten eventuell teilweise den verringerten Werten der Sprachheilschüler im KFT-Subtest geschuldet sein.

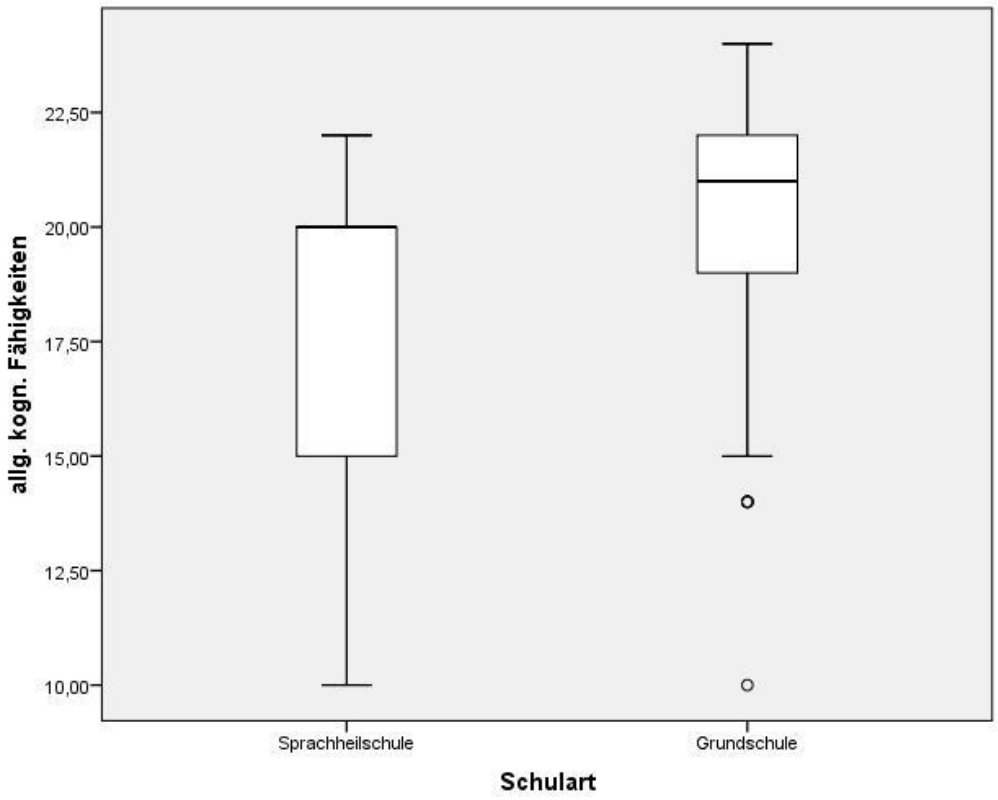

Abbildung 13: Validierung: Allgemeine kognitive Fähigkeiten in Abhängigkeit von der Schulart

Aus den deskriptiven Verteilungsanalysen der Feedbackresponsivitäten und der allgemeinen kognitiven Fähigkeiten folgt damit, dass der Unterschied in den Korrelationen möglicherweise weniger strukturellen Effekten der FR als eventuell vorhandenen (Boden-)effekten der allgemeinen kognitiven Fähigkeiten anzulasten ist. Zur abschließenden Klärung dieser Frage sind weitere Untersuchungen sind nötig. Ein weiterer Indikator für die mögliche prädiktive Validität der FR ist die Schullaufbahnempfehlung, die auf langfristige Erfolge/Misserfolge im schulischen Bereich hindeuten kann und der damit ein inkrementellen Beitrag zur Klärung der Frage nach den Zusammenhängen der Feedbackresponsivität zukommt.

Zur Schullaufbahnempfehlung

Um weitere Hinweise auf die Prädiktionskraft der FR zu erhalten, soll die Feedbackresponsivität als nächstes mit der Schullaufbahnempfehlung durch den jeweiligen Klassenlehrer in Zusammenhang gebracht werden. Tabelle 62 stellt die Befunde hierzu dar.

Da der Stichprobenumfang von n=9 gering ist, die Skala der Schullaufbahnempfehlung kein Ordinalskalenniveau erreicht und in ihren Ausprägungen starke Überlappungen aufweist, soll von einer statistischen Prüfung des Zusammenhangs abgesehen werden. Stattdessen werden die Einschätzungen der Lehrkräfte mit der Rangreihe der Feedbackresponsivität in Verbindung gebracht. Das Kind, welches am wenigsten auf die gegebenen Hilfestellungen anspricht, hat dabei Rangplatz 9.

Tabelle 62: Feedbackresponsivität und Schullaufbahnempfehlung in der Validierungsuntersuchung IIb

Rangplatz Feedbackresponsivität	**Schullaufbahnempfehlung**
1	Sonderschule für Sprachen mit Realschulzweig (Realschulabschluss)
2	SRH
3	Sprachheilschule
4	Sprachheilschule
5	Werkrealschule
6	Werkrealschule
7	Sprachheilschule
8	Werkrealschule
9	Werkrealschule/Förderschule

Die letzten drei Rangplätze liegen in ihren FR-Werten über dem arithmetischen Mittel der Stichprobe. Nur ein Kind liegt in seiner Feedbackresponsivität mehr als eine Standardabweichung vom Mittelwert entfernt: das Kind mit Rangplatz 1. Dieses Kind ist das einzige, dessen Schullaufbahnempfehlung explizit

Realschulniveau anspricht. Diese Befunde sprechen damit eher für die Validität der Feedbackresponsivität.

8.4.4 Interpretation

Zusammenfassend kann somit festgehalten werden, dass die Feedbackresponsivität erwartungsgemäß weder mit den erhobenen Lesemaßen noch mit den schulischen Leistungsmaßen, noch mit den Ängstlichkeitsmaßen korreliert.

Der nicht signifikante Zusammenhang der Feedbackresponsivität mit den allgemeinen kognitiven Fähigkeiten ist - insbesondere im Vergleich mit dem signifikanten Zusammenhang in der Population der Grundschüler - weniger eindeutig zu interpretieren. Möglich wäre ein in der Stichprobe der Sprachheilschüler eventuell auftretender Bodeneffekt in der KFT-Subskala, welche ursprünglich für die vierte Jahrgangsstufe einer Regelschule entwickelt wurde und hier unter anderem an einer dritten Jahrgangsstufe im Förderschulbereich eingesetzt wurde. Daneben deutet sich ein Zusammenhang der Feedbackresponsivität mit der Schullaufbahnempfehlung an. Das Kind, dessen Feedbackresponsivität herausragend gering war (wenige falsche Antworten im Zweitversuch), war das einzige Kind der Kohorte mit einer explizit auf Realschulniveau abzielenden Schulempfehlung. Das Kind, dessen relativer Anteil an falschen Antworten in der Stichprobe maximal war, erhielt als einziges Kind explizit die Schullaufbahnempfehlung „Werkrealschule/Förderschule“, wobei Förderschule sich als Begrifflichkeit nicht unbedingt auf Sprachheilschulen beschränken muss.

Warum das Lehrerurteil Lesen im Gegensatz zur Validierungsuntersuchung IIa einen stärkeren Zusammenhang mit der Feedbackresponsivität aufweist als die basale Lesefähigkeit, kann ebenfalls nicht abschließend geklärt werden. Vermutlich unterscheiden sich Pädagogen an Regel- und Förderschulen systematisch in den Maßstäben, nach denen sie Leistungen bewerten oder Lehrerurteile an Förderschulen schließen im Besonderen auch Urteile über

Entwicklungspotentiale im Bereich Lesen mit ein. Weitere Untersuchungen zur Klärung dieses Befunds sind nötig.

Die Korrelationen der Feedbackresponsivität in dieser spezifischen Population sind damit teilweise als hypothesenkonform zu sehen. Insgesamt kann jedoch nicht ausgeschlossen werden, dass die hier beobachteten Zusammenhänge Drittvariablen geschuldet sind, die nicht direkt erhoben wurden. Eine Erweiterung des erhobenen Variablenspektrums in Folgeuntersuchungen kann damit als explizites Ziel weiterer Untersuchungen festgehalten werden. Daneben verdienen die spezifischen methodischen Merkmale dieser Untersuchung besondere Beachtung und sollen nachfolgend nochmals aufgegriffen und diskutiert werden.

8.5 Diskussion

Nachfolgend sollen die Ergebnisse aller Validierungsstudien zusammengefasst bewertet und in einen größeren Kontext eingeordnet werden. Problematische methodische Aspekte sollen dabei ebenso berücksichtigt werden wie ein Ausblick auf weitere Projektschritte. Dabei werden zunächst generelle methodische Auffälligkeiten (Kapitel 8.5.1) und Besonderheiten der Untersuchungen an der Population mit spezifischem Förderbedarf thematisiert (Kapitel 8.5.2). Dem schließen sich Anmerkungen zur Untersuchung der dynamischen Komponente an (Kapitel 8.5.3). Nach einer abschließenden Bewertung der durchgeführten Validierungsuntersuchungen (Kapitel 8.5.4) soll der momentane Entwicklungsstand des Verfahrens anhand verschiedener Testgütekriterien umrissen und darauf aufbauend mögliche weitere Projektschritte skizziert werden (Kapitel 8.5.5).

8.5.1 Generelle methodische Anmerkungen

Einige ausgewählte methodische Aspekte der Validierungsuntersuchungen sollen nachfolgend in ihrer Bedeutung für die Validität der Ergebnisse diskutiert werden: die verminderten Reliabilitäten zweier Skalen in der Validierungsuntersuchung Ib, die systematische Varianzeinschränkung einzelner Skalen, die α-Fehler-Kumulierung und der Zusammenhang zwischen der allgemeinen Ängstlichkeit und der Testängstlichkeit.

Zu den verminderten Reliabilitäten in der Validierungsuntersuchung Ib

Generell weisen die Skalen in allen vier Validierungsuntersuchungen mindestens akzeptable Reliabilitäten auf. In der Validierungsuntersuchung Ib sind jedoch zwei Reliabilitäten auffallend. Die innere Konsistenz der Skala der allgemeinen kognitiven Fähigkeiten ist mit .665 (Cronbachs Alpha) eher niedrig und die innere Konsistenz der Skala der allgemeinen Ängstlichkeit ist mit .532 (Cronbachs Alpha) äußerst niedrig. Die Korrelation der Lesekompetenz mit den allgemeinen kognitiven Fähigkeiten liegt bei .444 und kann als mittelstark angesehen werden, die Korrelation der Lesekompetenz mit der allgemeinen Ängstlichkeit ist niedriger, sie liegt bei .155. Die Unterschiede

zu den entsprechenden Korrelationen in der Validierungsuntersuchung Ia könnten also auch daran liegen, dass die Tests in den unterschiedlichen Stichproben unterschiedlich reliabel sind. Diesem Umstand könnte mit einer Minderungskorrektur begegnet werden. Bei der Minderungskorrektur wird um die messfehlerbedingte Minderung der Korrelationen bereinigt (Pospeschill & Spinath, 2009), somit sollten sich die Korrelationen durch die Minderungskorrektur in ihrer Stärke erhöhen. Jedoch sprechen zwei Aspekte dafür, in diesem Zusammenhang auf eine Minderungskorrektur zu verzichten. Erstens entfernen sich durch sie die Korrelationen von den Korrelationen der Validierungsstichprobe Ia (Korrelation der Lesekompetenz mit den allgemeinen kognitiven Fähigkeiten: .404; Korrelation der Lesekompetenz mit allgemeiner Ängstlichkeit: −.082). Dies legt die Vermutung nahe, dass die Korrelationsunterschiede zwischen den Validierungsuntersuchungen nicht primär auf die verminderten Reliabilitäten in Validierungsuntersuchung Ib zurückzuführen sind. Daneben gibt die minderungskorrigierte Korrelation keinen wahren Zusammenhang an, sondern lediglich seine theoretische Obergrenze (Rentzsch & Schütz, 2009, S. 137). Sie ist damit in erster Linie von theoretischem Interesse, da sie „optimale“ Korrelationen ermittelt, die sich auf „optimale“ Messungen ohne Messfehler beziehen. Solche „optimalen“ Bedingungen liegen in der Realität jedoch nicht vor, die theoretisch mögliche Obergrenze kann nicht erreicht werden. Damit können die beobachteten empirischen Zusammenhänge durch eine Minderungskorrektur inhaltlich nicht besser aufgeklärt und verstanden werden.

Zur systematischen Varianzeinschränkung (*range restriction*)

Eine Unterschätzung der in der Validierung untersuchten Zusammenhänge kann daneben noch einem anderen technischen Aspekt geschuldet sein: der Abhängigkeit der Testwerte von der Grundgesamtheit (Pospeschill & Spinath, 2009, S. 189). Bei etlichen Variablen wurde mindestens ein Drittel der Skala nicht ausgeschöpft, bei den entsprechenden Variablen kann daher nicht ausgeschlossen werden, dass die gefundenen Korrelationen durch die systematische Varianzeinschränkungen dieser Skalen vermindert sind. Mehrere Gründe können dafür verantwortlich sein, dass bestimmte Skalen nicht ausgeschöpft werden. So können Selektionseffekte in diesem Zusammenhang

eine Rolle spielen: Leistungsstärkere und weniger testängstliche Kinder sind möglicherweise eher bereit, an einer wissenschaftlichen Untersuchung zur Lesekompetenz teilzunehmen. Daneben können Tendenzen zur sozialen Erwünschtheit dazu führen, dass schlechtere Schulnoten weniger häufig berichtet werden. Darüber hinaus sind insbesondere die Zensuren in Deutsch und Mathematik in der dritten und vierten Jahrgangsstufe nicht als gleichverteilt anzunehmen. Während höchstens ein Zehntel der Grundschüler in den Fächern Deutsch und Mathematik die Note „mangelhaft“ oder „ungenügend“ erhält, erhalten über ein Drittel die Zensuren „sehr gut“ und „gut“ (Krüsken, 2007, S. 49). Systematische Urteilseffekte wie beispielsweise Mildeeffekte auf Seiten der Lehrkräfte können in diesem Zusammenhang nicht komplett ausgeschlossen werden (vgl. Tent, 2006, S. 874) und können mögliche Ursachen für eine mangelnde Ausschöpfung der Notenskalen sein, welche die Streuung der empirisch ermittelten Daten systematisch einschränkt.

Die Unterschätzung der Korrelationen durch die systematische Varianzeinschränkung führt dazu, dass die (unter anderem von der Höhe der Korrelationen abhängige) Entscheidung für eine Signifikanz des Zusammenhangs eher konservativ ist. Dennoch korrelieren die FR und die Lesekompetenz meist signifikant mit den Variablen, bei denen auch ein Zusammenhang a priori erwartet wurde. Diese Befunde sprechen für die konvergenten Validitäten der Lesekompetenz. Dagegen ist der mögliche Einfluss der Korrelationsminderung auf die divergenten Validitäten kritischer zu bewerten. Insbesondere ist dies für die Validität der Feedbackresponsivität relevant, bei der hauptsächlich divergente Validitäten getestet wurden. Wenn hier die systematische Einschränkung der Varianz einzelner Variablen bedeutsame Effekte auf die jeweiligen Korrelationen hat, so müssen diese Korrelationen um den Einfluss der eingeschränkten Testwerte ihrer Grundgesamtheit korrigiert werden. Hierfür wäre eine Korrektur der Korrelationen nach Thorndike als besonders sinnvoll anzusehen (Wiberg & Sundström, 2009).

Gleichzeitig ist davon auszugehen, dass die systematische Varianzeinschränkung auch die Reliabilitätsschätzungen mindert (Pospeschill

& Spinath, 2009). Wie bereits ausgeführt, sind die empirisch ermittelten Reliabilitäten (von zwei Ausnahmen abgesehen) in allen vier Validierungsstichproben jedoch als hinreichend zu bewerten, was auf einen nicht bedeutsamen negativen Einfluss der systematischen Varianzeinschränkung auf die empirisch ermittelten Reliabilitäten hinweist. Damit sprechen die empirisch ermittelten Reliabilitäten eher weniger für eine im bedeutenden Maße eingeschränkte Streuung der empirischen Testwerte.

Darüber hinaus führen die systematisch eingeschränkten Testwerte zu unterschiedlichen Reliabilitäten für unterschiedliche Personen. Dies ist ein Problem, das insbesondere in den Extremwertbereichen bei Boden- und Deckeneffekten ausgeprägt ist (Schermelleh-Engel & Werner, 2012). Eine Möglichkeit, die Testergebnisse einzelner Probanden hinsichtlich ihrer Reliabilitäten zu verbessern, wäre beispielsweise das adaptive Testen, bei dem sich die Aufgabenschwierigkeiten den Fähigkeiten des Probanden anpassen (Pospeschill & Spinath, 2009). Eine adaptive Erweiterung des konstruierten dynamischen Lesekompetenztests wäre demnach als besonders vielversprechend einzustufen. Darüber hinaus sollte sie sich auch positiv auf motivationale Aspekte auswirken, da die Probanden weder unter- noch überfordert werden.

Daneben ist zu bedenken, dass die Wahrscheinlichkeit für eine systematische Varianzeinschränkung bei großem Stichprobenumfang verringert ist. So zeigt sich beispielsweise in der Validierungsuntersuchung Ia eine im Vergleich zu den anderen Studien besonders gute Ausschöpfung der Skalen. In Kapitel 8.5.2 wird die Bedeutung des Stichprobenumfangs, der insbesondere in den Untersuchungen der Sprachheilschüler verringert ist, nochmals thematisiert.

Zur α-Fehler-Kumulierung

Allen Validierungsuntersuchungen ist gemeinsam, dass mehrfach Korrelationen berechnet wurden, die sich aus denselben Daten speisen. Dies birgt die Gefahr der sogenannten α-Fehler-Kumulierung. Dieser kann durch ein adjustiertes Niveau des α-Fehlers entgegengewirkt werden, beispielsweise durch eine Bonferroni-Adjustierung (Eid, Gollwitzer & Schmitt, 2011, S. 400),

die auch bei Korrelationen angewandt werden kann (Bortz, 2005, S. 272). Jedoch ist sie bezüglich ihrer Entscheidung zur Signifikanz als tendenziell konservativ anzusehen, was insbesondere im Zusammenspiel mit dem sehr geringen Stichprobenumfängen der Validierungsuntersuchungen Ib und IIb problembehaftet sein könnte. Der Grund hierfür liegt in der Abhängigkeit der Ergebnisse des Signifikanztests vom Stichprobenumfang. Je höher der Stichprobenumfang, desto eher werden statistische Tests signifikant, je kleiner der Stichprobenumfang, desto eher werden statistische Tests nicht signifikant (wenn alle anderen Faktoren wie beispielsweise die Effektstärke konstant bleiben). Die Untersuchungen Ib und IIb haben einen sehr kleinen Stichprobenumfang und trotz teilweise vergleichbarer Effektstärken werden weniger Korrelationen signifikant als in den Untersuchungen Ia und IIa (vgl. Kapitel 8.5.2). Diese Tendenz, dass eigentlich bedeutsame Effektstärken keine Signifikanz erreichen, würde durch eine Bonferroni-Adjustierung noch verstärkt werden. Somit wird durch eine Adjustierung des α-Fehlers nicht automatisch ein Mehrwert hinsichtlich der Beantwortung der Validierungshypothesen erzielt.

Multivariate Analysemethoden wären in diesem Zusammenhang möglicherweise auch geeignet, dem Problem entgegen zu wirken. Mit ihnen lassen sich darüber hinaus auch die Beziehungen der externen Außenkriterien untereinander berücksichtigen. Dass von solchen Zusammenhängen ausgegangen werden kann, legen insbesondere die berichteten Korrelationen zwischen der Testängstlichkeit und der allgemeinen Ängstlichkeit nahe, auf die noch genauer eingegangen wird.

Zum Zusammenhang zwischen der allgemeinen Ängstlichkeit und der Testängstlichkeit

Die Korrelationen der allgemeinen Ängstlichkeit und der Testängstlichkeit sind in allen vier Validierungsuntersuchungen durchgehend hoch. Dies ist im Einklang mit den Befunden aus Kapitel 3.1.3.1, die Ähnlichkeiten zwischen Konzepten nahe legen. Außerdem stammen beide Skalen aus demselben Instrument (vgl. Kapitel 8.1.2.2), was eine Ähnlichkeit der gemessenen Konstrukte begünstigen kann.

In der Validierungsuntersuchung Ia korrelieren die beiden Ängstlichkeitsmaße mit r=.631 miteinander. Trotz dieser hohen Korrelation zeigen sich zwischen der allgemeinen Ängstlichkeit und der Testängstlichkeit Unterschiede in den Zusammenhängen mit der Lesekompetenz. Die Korrelation der Lesekompetenz mit der Testängstlichkeit wird signifikant, die Korrelation der Lesekompetenz mit der allgemeinen Ängstlichkeit nicht. Ein ähnliches Bild ergibt sich bei der Validierungsuntersuchung Ib. Die Testängstlichkeit und die allgemeine Ängstlichkeit hängen mit r=.794 signifikant miteinander zusammen. Die Lesekompetenz korreliert mit der Testängstlichkeit zu r=.051 und mit der allgemeinen Ängstlichkeit zu r=.155. Würde es sich bei Testängstlichkeit und allgemeiner Ängstlichkeit im Prinzip um dasselbe Konstrukt handeln, so müssten beide ähnliche Korrelationen mit der Lesekompetenz aufweisen. Dies ist jedoch in der Validierungsuntersuchung Ib nicht der Fall.

Ein Vergleich der Korrelationen mit dem von Lenhard und Lenhard (2014) entwickeltem Webinterface, welches auf dem bei Eid et al. (2011) vorgeschlagenem Vorgehen basiert, liefert folgendes Bild: Die Korrelation der Lesekompetenz mit der Testängstlichkeit unterscheidet sich nicht signifikant von der Korrelation der Lesekompetenz mit der allgemeinen Ängstlichkeit (z=1.332, p=.091). Dieser von der Validierungsuntersuchung Ia abweichende Befund ist jedoch mit Blick auf die verminderte Reliabilität der Skala der allgemeinen Ängstlichkeit und insbesondere mit Blick auf den verminderten Stichprobenumfang der Untersuchung der Sprachheilschüler mit Vorsicht zu interpretieren.

Insgesamt kann bezüglich der Lesekompetenz festgehalten werden, dass sie zwischen der allgemeinen Ängstlichkeit und der Testängstlichkeit diskriminiert. Die Korrelationen mit diesen beiden emotionalen Konstrukten sind unterschiedlich hoch, obwohl sich die beiden Ängstlichkeitsmaße ähnlich sind. Methodische Gründe könnten erklären, warum sich die Korrelationen der beiden Ängstlichkeitsmaße mit der Lesekompetenz nicht signifikant voneinander unterscheiden. Beide Indikatoren, die allgemeine Ängstlichkeit und die Testängstlichkeit, haben somit ihre Berechtigung. Sie bringen für die

zu klärende Fragestellung inkrementellen Nutzen und es erscheint wenig zielführend, sie aggregiert zu betrachten.

Bei den Untersuchungen zur dynamischen Komponente des Lesekompetenztests liegt in Validierungsuntersuchung IIb der Zusammenhang zwischen der Testängstlichkeit und der allgemeinen Ängstlichkeit bei r=.834. Es zeigen sich durchweg positive Korrelationen der FR mit der allgemeinen Ängstlichkeit (linear: r=.261; monoton: r= .299) und mit der Testängstlichkeit (linear: r=.278; monoton: r=.275). Sowohl die linearen als auch die monotonen Korrelationen können dabei als relativ homogen angesehen werden, eine inferenzstatistische Absicherung der Unterschiede erscheint auf Grund des besonders geringen Stichprobenumfangs jedoch als wenig vielversprechend.

In Validierungsuntersuchung IIa liegt der Zusammenhang zwischen Testängstlichkeit und allgemeiner Ängstlichkeit bei r=.729. Es zeigt sich hier bereits an den Vorzeichen der Korrelationen, dass die Korrelation der FR mit der allgemeinen Ängstlichkeit (linear: r=−.144; monoton: r=−.099) der Korrelation der FR mit der Testängstlichkeit (linear: r=.130; monoton: r= .190) weniger ähnelt als dies in der Population der Sprachheilschüler der Fall war. Die FR hängt positiv mit Testängstlichkeit und negativ mit allgemeiner Ängstlichkeit zusammen. Dabei ist insbesondere die Spearman-Korrelation der FR mit der allgemeinen Ängstlichkeit relativ nahe 0.

Diese unterschiedlichen Befunde können spezifische Populationsunterschiede wiederspiegeln. Möglicherweise ist das Verhalten von Sprachheilschülern im zweiten Versuch stärker von Ängstlichkeit dominiert, wobei zwischen allgemeiner Ängstlichkeit und Testängstlichkeit nur bedingt diskriminiert werden kann. Nicht auszuschließen sind aber auch mögliche methodische Unterschiede zwischen den Validierungsuntersuchungen IIa und IIb, die es noch weiter zu ergründen gilt. Systematische methodische Besonderheiten der Validierungsuntersuchungen an Sprachheilschülern sollen nachfolgend noch genauer ausgeführt werden.

8.5.2 Anmerkungen zu den Befunden der Validierungsuntersuchungen Ib und IIb an Schülern mit spezifischem Förderbedarf

Zum Stichprobenumfang

Eine besondere Schwachstelle der Erhebung im Förderschulbereich ist die geringe Anzahl an Probanden. Sie ist neben vereinzelt auftretenden technischen Problemen auch der Tatsache geschuldet, dass jedes Kind in einer Einzeltestung von etwa zwei Schulstunden erhoben wurde. Während die Testsituation der Einzeltestung in Hinblick auf die Population und ihre spezifischen Bedürfnisse als gerechtfertigt angesehen werden kann, so ist gleichzeitig der geringe Stichprobenumfang in Hinblick auf sämtliche Befunde problematisch, die in diesen Untersuchungen präsentiert wurden.

Ein positiver Aspekt bezüglich der Validität der Ergebnisse der Einzeltestung wird ersichtlich, wenn die fehlenden Werte durch unmotivierte Probanden betrachtet werden. Kinder, die sich „durchklicken“ und eine generell gering ausgeprägte Motivation an den Tag legen, finden sich ausschließlich in der Population der Grundschüler. Die Erhebungen in den Grundschulen wurden gruppenweise durchgeführt, bei den Sprachheilschülern waren die Testleiter meist direkt neben dem Probanden. Damit waren die Sprachheilschüler einer höheren sozialen Kontrolle ausgesetzt, welche sich offensichtlich positiv auf die Arbeitshaltung auswirkte. Dies ist in Einklang mit Befunden von Golke et al. (2015), die auf die Bedeutung der Motivation der Probanden hinweisen.

Dem gegenüber stehen die negativen Auswirkungen des verringerten Stichprobenumfangs, welche eine verringerte statistische *Power* nach sich zieht. Wahre Effekte können durch die Signifikanzprüfung damit weniger wahrscheinlich als wahre Effekte erkannt werden und die Wahrscheinlichkeit, dass ein statistisch signifikantes Ergebnis einen wahren Effekt wiedergibt, ist verringert (Button et al., 2013). Dies könnte beispielsweise auf die in der Validierungsuntersuchung Ib ermittelte Korrelation zwischen den allgemeinen kognitiven Fähigkeiten und der Lesekompetenz zutreffen (vgl. Kapitel 8.2.3).

Für die Bewertung der Validierungsstudien ist insbesondere von Bedeutung, dass nicht signifikanten Korrelationen dennoch ein wahrer Effekt zu Grunde liegen kann.

Tabelle 63: Ausgewählte Stichprobenumfänge und die daraus abgeleiteten minimalsten Korrelationen, welche Signifikanz erreichen

Stichprobenumfang	**Minimalste Korrelation**
3	.999
4	.974
5	.935
9	.790
10	.761
14	.632

Zur Veranschaulichung dieser Problematik sind in Tabelle 63 die mit *G*Power* 3.0.10 (Faul, Buchner, Erdfelder & Lang, 2008) berechneten minimalsten Korrelationen ersichtlich, bei denen statistische Testungen unter Annahme eines bestimmten, fest vorgegebenen Stichprobenumfang signifikant werden (zweiseitige Testung, H_0: ρ=0, α=.05, β=.80). So kann beispielsweise bei einer Stichprobengröße von 14 die in Kapitel 8.4.3 berichtete Korrelation der Lesekompetenz mit den basalen Lesefähigkeiten signifikant werden, da sie mit .732 größer als .632 ist. Die Korrelation der Lesekompetenz mit den allgemeinen kognitiven Fähigkeiten liegt dagegen bei .444. Sie müsste jedoch bei diesem Stichprobenumfang mindestens .632 erreichen, um signifikant werden zu können. Beim Vergleich der Korrelationen der Validierungsuntersuchungen an Grundschülern mit ihrem jeweiligen Pendant an Sprachheilschülern wird daher ersichtlich, dass es im Sinne der zu beantwortenden Fragestellungen nicht zielführend ist, der Signifikanzprüfung allzu starkes Gewicht zu geben. Vielmehr sollen auch die tatsächlich beobachteten Korrelationen in ihrer Größe und Richtung Berücksichtigung finden. Die Relevanz des verringerten Stichprobenumfangs zeigt sich beispielsweise in der signifikanten Korrelation der Lesekompetenz mit den

allgemeinen kognitiven Fähigkeiten in der Validierungsuntersuchung Ia. Sie ist mit r=.404 nicht größer als die entsprechende Korrelation in der Untersuchung der Sprachheilschüler (r=.444), die nicht zufallskritisch abgesichert werden kann.

Vergleich mit den Validierungsuntersuchungen Ia und IIa an Schülern der Grundschulen

Die beobachteten Korrelationen können überdies zu den in den Validierungsstudien Ia und IIa gefundenen Korrelationen (Kapitel 8.1.3 und Kapitel 8.3.3) in Bezug gesetzt werden. Von einer statistischen Prüfung der Korrelationsdifferenzen soll allerdings abgesehen werden. Der Grund hierfür ist, dass in den Förderschulpopulationen die Fallzahlen gering sind und diese damit große Konfidenzintervalle und als Folge dessen auch höhere Überlappungswahrscheinlichkeiten der Konfidenzintervalle bedingen. Bei Fallzahlen im einstelligen Bereich wird auf eine Bewertung verzichtet, da hier der individuelle Einfluss der einzelnen Versuchsperson einen zu großen Einfluss auf die jeweilige Korrelation hat.

In Hinblick auf die Lesekompetenz fällt in der Population der Sprachheilschüler auf, dass die Korrelationen mit den Lesemaßen einen größeren Betrag aufweisen als in der Gruppe der Grundschüler, während die Korrelation zwischen der FR und der basalen Lesefähigkeit erwartungsgemäß auch in der Förderschule gering bleibt.

Die Korrelation mit den allgemeinen kognitiven Fähigkeiten fällt bei den Sprachheilschülern für die Feedbackresponsivität in ihrem Betrag geringer aus als die entsprechenden Korrelationen in der Validierungsstichprobe der Grundschüler. Allgemeine kognitive Fähigkeiten scheinen damit in Hinblick auf die FR in der Population der Sprachheilschüler eine geringere Rolle zu spielen. Es kann nicht ausgeschlossen werden, dass eine im Förderschulbereich eher vorliegende systematische Varianzeinschränkung im allgemeinen kognitiven Fähigkeitsbereich hierbei eine Rolle spielt. In Hinblick auf die deskriptiven Kennwerte der Verteilung lässt sich bei den Förderschülern eine

verringerte Performanz im KFT-Subtest feststellen, die jedoch aus bereits dargelegten Gründen nicht statistisch abgesichert werden soll.

Zur Erfassung der allgemeinen kognitiven Fähigkeiten muss an dieser Stelle jedoch noch kritisch angemerkt werden, dass die Messung mit einer einzelnen Subskala des Kognitiven Fähigkeitstests (KFT) erfolgte und diese Subskala lediglich für die vierte Klassenstufe entwickelt und normiert wurde. Damit sind insbesondere die Ergebnisse der dritten Jahrgangsstufe und die Ergebnisse der Population mit spezifischem Förderbedarf möglicherweise in ihrer Aussagekraft eingeschränkt.

Demgegenüber sind die verwendeten Notenskalen als von dieser Problematik nicht betroffene Messinstrumente der akademischen Leistung anzusehen. Im Gegensatz zu den mit Grundschülern durchgeführten Validierungsstudien wird die sechsstufige Notenskala bei der Population der Sprachheilschüler ausgeschöpft. Systematische Varianzeinschränkungen spielen damit z. B. beim Lehrerurteil Lesen möglicherweise eine eher geringere Rolle und ergänzen somit gezielt die Befunde zu den allgemeinen kognitiven Fähigkeiten.

Im Gegensatz zu den Grundschülern scheinen im Förderschulbereich insbesondere ängstliche Kinder (hohe Werte in beiden Ängstlichkeitsmaßen) in der Tendenz besonders häufig nicht vom gegebenen Feedback zu profitieren und damit im Zweitversuch falsch zu antworten, wenngleich eine statistische Absicherung der Korrelationsunterschiede aus oben genannten Gründen nicht zielführend ist. Die Feedbackresponsivität scheint auch in höherem Maße mit der Mathematiknote und dem Lehrerurteil Lesen übereinzustimmen, jedoch können diese Tendenzen aus methodischen Gründen nicht zufallskritisch abgesichert werden. Auch der Einfluss der systematischen Varianzeinschränkung kann hier nicht komplett ausgeschlossen werden. Daneben könnten auch inhaltliche Gründe dafür verantwortlich sein, dass sich eine verringerte Lernfähigkeit im Förderschulbereich im besonderen Maße negativ in der Notengebung niederschlägt. Systematische Unterschiede in der Benotung zwischen Lehrern an Förder- und Regelschulen könnten hier eine Rolle spielen, beispielsweise eine stärkere Berücksichtigung des individuellen

Potentials des Schülers an Sprachheilschulen. Jedoch sind die Fallzahlen gering und es lässt sich nicht ausschließen, dass die hier beobachteten Zusammenhänge mit den Komponenten des dynamischen Lesekompetenztests populationsabhängig sein können.

Empirisch unterscheidet sich die Ausprägung der Feedbackresponsivität nicht zwischen Förder- und Regelschulen, eine Aussage, die sich für die allgemeine kognitive Leistung gemessen durch die KFT-Subskala in dieser Form nicht treffen lässt (vgl. Kapitel 8.4.3). Dieser Befund sollte nochmals an einer größeren Stichprobe repliziert werden. Falls er sich bestätigt, so wäre das im Einklang mit den theoretischen Überlegungen zur dynamischen Testung (Kapitel 2.7) und würde nicht gegen die Validität des hier erhobenen Konstrukts der Feedbackresponsivität sprechen.

Es bleibt abzuwarten, ob das hier berichtete Korrelationsmuster in seiner Richtung und seinem Betrag stabil bleibt, wenn der Stichprobenumfang vergrößert wird. Bis dies abschließend geklärt ist, sind die in diesen Erhebungen ermittelten Befunde als vorläufige Ergebnisse der Validierungsuntersuchungen Ib und IIb anzusehen.

Zwischenfazit

Die bisher diskutierten Befunde in der Population der Schüler mit spezifischem Förderbedarf deuten insgesamt darauf hin, dass die Lesekompetenz wohl als ein wichtiges Korrelat der externen Lesemaße und der allgemeinen kognitiven Fähigkeiten, jedoch nicht als ein bedeutsames Korrelat verschiedener Ängstlichkeitsmaße angesehen werden kann.

Die Feedbackresponsivität scheint ebenfalls kein starkes Korrelat dieser Ängstlichkeitsmaße zu sein. Ob die Feedbackresponsivität als ein bedeutsames Korrelat der kognitiven und auf schulischen Erfolg abzielenden Maße anzusehen ist, lässt sich noch nicht abschließend beantworten.

8.5.3 Anmerkungen zu den Befunden der Validierungsuntersuchungen IIa und IIb der dynamischen Komponente

Psychometrische Gesichtspunkte der Feedbackresponsivität

Zur Umsetzung der Feedbackresponsivität sei an dieser Stelle noch auf einige der in Kapitel 2.6 dargelegte Problembereiche des dynamischen Assessments hingewiesen, die nochmals aufgegriffen und diskutiert werden sollen. Generell war es Ziel der vorliegenden Arbeit, bei der Umsetzung der Feedbackresponsivität möglichst hohen psychometrischen Standards zu genügen.

Ein wesentlicher Schwachpunkt der Erfassung der dynamischen Komponente der dynamischen Testung ist, dass die Skala allgemein noch nicht hinreichend gut verstanden ist (vgl. Kapitel 2.6). Ihr Wesen abschließend zu klären war weder Ziel noch Anspruch der vorliegenden Arbeit. Vielmehr sollte eine Operationalisierung der dynamischen Komponente entwickelt werden, welche dem Zweck des Projekts dienlich war und in den hier aufgeführten Studien eine valide Messung erlaubte. Dieses Ziel konnte mit dem Indikator der Feedbackresponsivität erfüllt werden.

Hieran knüpft eine weitere psychometrische Konsequenz für die durchgeführten Studien an. Da die Skala der dynamischen Maße noch nicht ausreichend gut verstanden ist, erscheint es bei der erstmaligen Anwendung eines neuen Indikators der dynamischen Testkomponente sinnvoll, möglichst wenig Vorannahmen über die Art und Weise zu treffen, wie dieser Indikator mit externen Kriterien zusammenhängt. Insbesondere kann eine Linearität a priori nicht postuliert werden. Um diesen Umstand angemessen zu berücksichtigen, wurde daher die Monotonie der Zusammenhänge mit Hilfe der Korrelationskoeffizienten nach Spearman gesondert betrachtet. Die Monotonieannahme ist leichter als die Linearitätsannahme zu erfüllen, da die Linearität einer Beziehung als eine hinreichende Voraussetzung für die Monotonie angesehen werden kann und damit ohne Monotonie keine Linearität

vorliegen kann. Insgesamt legen die Befunde der Validierungsstichprobe IIa nahe, dass sich bei den untersuchten Zusammenhängen die Stärke der Monotonie und die Stärke der Linearität insgesamt nicht auffallend unterscheiden. Damit scheint der monotone Zusammenhang zwischen den jeweiligen Variablen relativ linear zu sein. Eine Ausnahme bilden lediglich das Lehrerurteil Lesen und die basale Lesefähigkeit, was jedoch auch an Spezifika dieser Skalen liegen kann. Stärkere Abweichungen zwischen den Korrelationsmaßen finden sich auch in der Population der Sprachheilschüler. Hier fallen insbesondere die Zusammenhänge mit der basalen Lesefähigkeit und die Zusammenhänge mit der Schulnote in Mathematik auf. Es ist nicht auszuschließen, dass diese Unterschiede sich bei größeren Stichprobenumfängen (unter Annahme der Stabilität der Werte) in den Ergebnissen der Signifikanztests niederschlagen würden. Ihre exakten Zusammenhänge zur Feedbackresponsivität liegen jedoch außerhalb des zentralen Fokus dieser Arbeit.

Ein weiteres Problemfeld betrifft die Messgenauigkeit. Die Reliabilitäten der Differenzen zwischen Prä- und Posttest sind höher, wenn die Korrelation zwischen Prä- und Posttest abnimmt, d. h. wenn mindestens eine dieser beiden Messungen weniger reliabel wird (vgl. Kapitel 2.6). Es wurde in der vorliegenden Arbeit versucht, dieses Problem zu verringern, indem Differenzen nicht derart explizit im Vordergrund der Messung standen, wie beispielsweise beim Ansatz von Budoff (Kapitel 2.5). Vielmehr steht der Anteil an Aufgaben im Vordergrund, bei denen es keinen Unterschied, also keine Differenzen gab. Dass damit keine *gainer* im Sinne Budoffs zu ermitteln sind, ist nicht unbedingt als Nachteil zu betrachten. Vielmehr stellt sich die Frage, inwieweit die Dichotomisierung in *gainer* und *non-gainer* der Komplexität des Sachverhaltes gerecht wird. So können sich beispielsweise zwei als *gainer* bezeichnete Probanden aus derselben Gruppe unähnlicher sein als zwei Probanden, von denen einer als *gainer* und einer als *non-gainer* klassifiziert wird. Insbesondere in Hinblick auf die Zuteilung zu pädagogischen Fördermaßnahmen kann sich diese Dichotomisierung als problembehaftet erweisen. Dem gegenüber kann mit dem in diesem Projekt verwendeten Ansatz ein Informationsverlust durch Dichotomisierung verhindert werden.

Inwieweit die verminderte Ökonomie gegenüber statischen Verfahren (vgl. Kapitel 2.6) durch einen höheren diagnostischen Mehrwert des dynamischen Assessments gerechtfertigt erscheint, lässt sich nach momentaner Kenntnislage noch nicht abschließend klären. Jedoch ist der in dieser Arbeit verwendete *train-within-test*-Ansatz der dynamischen Testung in Hinblick auf die Durchführungsökonomie dem *test-train-test*-Design überlegen. Des Weiteren ist festzuhalten, dass der in der hier dargestellten Testentwicklung verwendete Ansatz die dynamische Komponente ökonomischer realisierte als dies bei den in Kapitel 2.5 beschriebenen Ansätzen von Campione und Brown, Guthke, sowie Carlson und Wiedl der Fall war. Der Grund liegt darin, dass nur ein Feedback pro Item gegeben wurde. Das führte zu einer Ökonomisierung der Testentwicklung und Testanwendung. Der Nachteil des Verfahrens liegt darin, dass weniger stark im Ausmaß differenziert wird, in welchem die gegebene Rückmeldung dem Kind hilft bzw. nicht hilft. Dieses Ausmaß kann beispielsweise durch die Anzahl an benötigten Rückmeldungen bis zur richtigen Antwort operationalisiert werden. Möglicherweise hätte ein solcher Ansatz insbesondere bei der Population mit spezifischem Förderbedarf einen diagnostischen Mehrwert gebracht. Um den Gewinn an Informationen zu maximieren, könnte bei *test-train-test*-Testungen auch alternative Ansätze in Betracht gezogen werden. So kann beispielsweise das letzte Drittel einer Testung als Posttest und die ersten zwei Drittel als Prätest aufgefasst werden (Guthke & Wiedl, 1996, S. 116). Für diesen Ansatz müsste jedoch der Lesekompetenztest derart verändert werden, dass er nicht mehr diese breite Kombination an Aufgaben- und Informationsarten abdeckt. Es muss sorgfältig abgewogen werden, ob die so verminderte Breite des Lesekompetenzkonstrukts durch einen diagnostischen Mehrwert bei der Erhebung der dynamischen Komponente aufgewogen werden kann.

Zur Generalisierbarkeit der Befunde

Die Befunde können in zweierlei Hinsicht auf ihre Generalisierbarkeit hin bewertet werden. Zum einen stellt sich die Frage, ob die hier gefundenen Ergebnisse auf andere Kulturtechniken übertragbar sind. Diese Frage lässt sich nicht abschließend klären, obgleich die Feedbackresponsivität als von der Lesekompetenz unabhängig angenommen werden kann. Sie sollte damit in

ähnlicher Form ihre Validität und Wirksamkeit zeigen, wenn sie in anderen akademisch-schulischen Kompetenzbereichen Anwendung findet.

Zum anderen kann die Generalisierbarkeit auf andere Populationen betrachtet werden, welche in ihrer Beschaffenheit (Schüler der Klassenstufen drei und vier) den hier untersuchten Populationen ähneln. Während die großen Stichprobenumfänge in der Pilotierung und insbesondere in der Validierung für eine Generalisierbarkeit der dort dargelegten Befunde und damit für eine Testkonstruktion sprechen, die valide Messungen begünstigt, so gibt es doch zwei Einschränkungen in Hinblick auf die Übertragbarkeit der Validitätsannahmen der Testung. Die erhobenen Stichproben sind zum einen lokal massiert. Sie wurden durchgehend nur in Baden-Württemberg erhoben. Damit unterliegen alle getesteten Schüler demselben Lehrplan. Darüber hinaus wurde innerhalb Baden-Württembergs überwiegend nur in einem begrenzten regionalen Umfeld erhoben. Die aus diesen Erhebungen stammenden Daten sind daher nicht automatisch als repräsentativ für den gesamten deutschsprachigen Raum anzusehen.

Ein weiterer die Generalisierbarkeit einschränkender Aspekt könnte die Freiwilligkeit der Teilnahme sein. Neben den Einverständniserklärungen der Eltern war auch der Wille des Kindes für die Teilnahme an der Untersuchung ausschlaggebend. Es ist anzunehmen, dass die Einverständniserklärung der Eltern und die Bereitwilligkeit der Kinder zur Teilnahme nicht unabhängig von sozialen Faktoren wie dem Bildungsniveau sind. Hinzu kommt, dass sowohl die Motivation der Kinder als auch die Aufgeschlossenheit der Eltern der Untersuchung gegenüber möglicherweise mit den Fähigkeiten und schulischen Leistungen der Kinder in Zusammenhang stehen. Leistungsstärkere Kinder sind möglicherweise mehr daran interessiert, an einer von Mitarbeitern einer Hochschule durchgeführten Untersuchung zur Lesekompetenz teilzunehmen. Damit würde sich erklären lassen, warum die Notenskalen bei den erhobenen Grundschülern nicht ausgeschöpft und somit Schüler mit guten Noten überrepräsentiert sind. Auch könnten tendenziell weniger prüfungsängstliche Schüler eher bereit sein, an einer Testung zu partizipieren. Daneben können

motivationale Faktoren auf Seiten der Kinder ihre Entscheidung beeinflussen, an der Untersuchung zu partizipieren.

Des Weiteren stellt sich die Frage, ob die Freiwilligkeit der Teilnahme der Schulen zu einer Verzerrung der Befunde führt. Möglich wäre beispielsweise, dass motivationale Faktoren (beispielsweise Engagement) auf Seiten der Schulleitung und des Lehrerkollegiums sich auf die Bereitschaft zur Teilnahme an einer psychologischen Testung auswirken und diese Faktoren sich auch in der Qualität des Unterrichts positiv niederschlagen.

Auch die Ausstattung mit PC-Räumen, welche eine Grundvoraussetzung für die Durchführung des dynamischen Lesekompetenztests an Grundschulen war, kann als nicht zufällig variierend eingeschätzt werden und es kann nicht ausgeschlossen werden, dass Schulen mit besserer Ausstattung generell über mehr Ressourcen verfügen, was sich in den Leistungen der erhobenen Kinder wiederspiegeln kann.

Insgesamt sprechen diese Punkte jedoch nicht dafür, eine Generalisierbarkeit der hier ermittelten empirischen Befunde komplett in Frage zu stellen. Vielmehr bieten sie Ansatzpunkte für weitere Forschung.

Ausblick

Eine Ausweitung der Untersuchung auf Grundschulen außerhalb Baden-Württembergs würde einen wertvollen Beitrag dazu leisten, den dynamischen Lesekompetenztest in seiner Gültigkeit noch besser generalisieren zu können. Wünschenswert wäre dabei, auch Kinder zu akquirieren, die bislang wenig in den erhobenen Stichproben repräsentiert sind, beispielsweise Grundschüler mit tendenziell schlechteren Schulnoten.

Daneben würde auch dem kleinen Stichprobenumfang an Sprachheilschülern und der damit einhergehenden methodischen und inhaltlichen Problematik in der Ergebnisdarstellung und deren Interpretation durch weitere Erhebungen entgegengewirkt werden. Sie könnten darüber hinaus mit einer Erweiterung des erhobenen Variablenspektrums verknüpft werden und damit das Verständnis

für die bislang berichteten Zusammenhänge verstärken. Damit könnte auch dem in Kapitel 2.6 genannten Kritikpunkt begegnet werden, wonach das dynamische Assessment als Forschungsgegenstand zu wenig Berücksichtigung findet.

8.5.4 Abschließende Bewertung der Validierungsstudien

Das Ziel der Validierungsstudie war, die im dynamischen Lesekompetenztest erfassten Konstrukte in Beziehung zu externen Außenkriterien zu setzen und an Hand ihrer Zusammenhänge mit diesen spezifischen Außenkriterien abschätzen zu können, ob der Test die Konstrukte erfasst, die er messen soll. Dies konnte umgesetzt werden.

Die Ergebnisse der Validierungsstudien sind sowohl für die Lesekompetenz als auch für die Responsivität auf das gegebene Feedback als positiv zu bewerten. Für beide Konstrukte finden sich Korrelationsmuster, welche sich mit den jeweiligen theoretischen Annahmen decken. Daneben können beide Konstrukte als hinreichend voneinander verschieden angesehen werden. Somit sind sowohl die Lesekompetenz als auch die Feedbackresponsivität mit dem in diesem Projekt entwickelten Instrument valide erhoben.

Es besteht darüber hinaus auch kein Spannungsverhältnis zum Befund M.1. der Metaanalyse (Kapitel 4.3), wonach die dynamische Komponente des dynamischen Assessments mit der statischen Lesekompetenz signifikant positiv korreliert, da für die Validierung die dynamische Komponente explizit um die Lesekompetenz bereinigt wurde.

Daneben sprechen auch die empirisch ermittelten Testreliabilitäten nicht gegen die Validität des Instruments. Die internen Konsistenzen von .706 bis .817 sind ein Indikator für eine ausreichende Homogenität der Items. Die im Vergleich zum Salzburger Lese-Screening verminderten Reliabilitäten lassen sich inhaltlich begründen und sind damit nicht negativ zu bewerten. Durch die Kombination verschiedener Aufgaben- und Informationsarten (siehe Kapitel 5.2) ist der in diesem Projekt konstruierte Lesekompetenztest als wesentlich heterogener anzusehen als das Salzburger Lese-Screening. Skalen, die ein gewisses Maß an Heterogenität abdecken sollen, haben eine verringerte Messgenauigkeit gegenüber Skalen, welche homogene psychologische Konstrukte erfragen (Schmidt-Atzert & Amelang, 2012, S. 49).

Die Umsetzung der Validierungsuntersuchung in zwei voneinander unabhängige Validierungsstudien mit unterschiedlichen Stichproben mag auf den ersten Blick unnötig erscheinen. Jedoch wäre eine einzige Validierungsstudie mit ein und derselben Stichprobe aus methodischer Sicht möglicherweise problematisch gewesen (Stichprobenkonfundierung). Eine Absicherung der Validität der Lesekompetenzkomponente wäre nur in Abhängigkeit von der dynamischen Komponente möglich gewesen, ein Messwiederholungseffekt hätte nicht ausgeschlossen werden können. Dagegen werfen die unterschiedlich großen Stichprobenumfänge der beiden Studien Ia und IIa keine methodischen Probleme in Hinblick auf die Fragestellungen auf, die in der Validierung beantwortet werden.

An dieser Stelle sei noch eine kritische Anmerkung gemacht: ein stichhaltiger Beweis für die Validität ist allein mit erwartungskonformen Korrelationen noch nicht gegeben. Aber sie sind erste Hinweise darauf, dass das erfasste Konstrukt mit Lesen in Zusammenhang steht bzw. im Falle der Feedbackresponsivität eine gewisse Überlappung mit kognitiven Fähigkeiten aufweist, deren Zustandekommen nicht befriedigend durch Zufall erklärbar ist.

Abschließend lässt sich festhalten, dass die beiden Validierungsstudien insgesamt zeigen konnten, dass das in diesem Projekt entwickelte Instrument des dynamischen Lesekompetenztests für eine sinnvolle Anwendbarkeit in der pädagogisch-psychologischen Praxis vorläufig als hinreichend gut legitimiert angesehen werden kann.

8.5.5 Der aktuelle Entwicklungsstand des dynamischen Lesekompetenztests im Kontext ausgewählter Testgütekriterien

Die Validierung des dynamischen Lesekompetenztests legt insgesamt nahe, dass sowohl die statische Erhebung der Lesekompetenz als auch die Umsetzung der Feedbackresponsivität hinreichend gut funktionieren. Ein neues dynamisches Testverfahren im Bereich der Lesekompetenz für Kinder der dritten und vierten Jahrgansstufe im deutschsprachigen Raum kann somit als vorläufig bewährt angesehen werden. Jedoch sind die bislang durchgeführten Untersuchungen für eine abschließende Testbewertung noch nicht ausreichend. Weitere Untersuchungen sind nötig, um die bisherigen Befunde zu replizieren und die aktuelle Testversion gegebenenfalls weiterzuentwickeln. Die übergeordnete Zielsetzung ist hierbei die Qualität des Testverfahrens, die sich anhand von verschiedenen Testgütekriterien beschreiben lässt.

Für eine erste Bewertung des konstruierten dynamischen Lesekompetenztests bietet sich eine Einordnung des Testverfahrens anhand psychologischer Testgütekriterien an. Hierbei sind in erster Linie die Hauptgütekriterien der Objektivität, der Reliabilität und der Validität zu nennen.

Überlegungen zur Objektivität wurden bereits in Kapitel 2 und Kapitel 5.1 angestellt. Durch die Umsetzung einer computeradministrierten Testung kann die Durchführungs- und Auswertungsobjektivität als gut bezeichnet werden. Mit Blick auf die Interpretationsobjektivität sind jedoch noch Maßnahmen erforderlich. Bislang können die Testergebnisse eines Probanden nur im Sinne von „hoch“ und „niedrig“ unterschieden werden, beispielsweise kann ein hoher Anteil richtiger Antworten im Erstversuch als hohe Lesekompetenz aufgefasst werden. Dies lässt sich jedoch noch weiter präzisieren. So wären beispielsweise deskriptive Skalenmerkmale eine Hilfestellung, um einzelne Testwerte in Bezug zu relevanten Vergleichsgruppen (z. B. bezüglich der Klassenstufe) setzen zu können. Dabei muss der eventuelle Bezug zu relevanten Vergleichsgruppen zur übergeordneten Zielstellung des Tests in

seinem anvisierten Einsatzbereich und den sich daraus ableitenden spezifischen Anforderungen passen (vgl. Goldhammer & Hartig, 2012; Bortz & Döring, 2006).

Durch eine Normierung des Tests kann daneben noch dem Testgütekriterium der Fairness Rechnung getragen werden. Obgleich die Befunde der Validierungsstudien Performanzunterschiede beispielsweise zwischen Mädchen und Jungen sowie zwischen Schülern der dritten und vierten Jahrgangsstufe nahelegen, werden diese Gruppen momentan noch aggregiert betrachtet. Die Bewertung der Testwerte ist damit momentan noch diskriminierend, da sie die Performanz bestimmter Gruppen systematisch unterschätzen und sie damit benachteiligen. So kann beispielsweise ein Schüler der dritten Klasse über eine überdurchschnittliche Lesekompetenz verfügen. Jedoch kann er nicht notwendigerweise einen überdurchschnittlichen Testwert in einer Vergleichsgruppe erreichen, die auch Schüler der vierten Klassenstufe umfasst, die gegenüber Schülern der dritten Klassenstufe eine tendenziell höhere Lesekompetenz aufweisen. Ausgehend von den Befunden der Testvalidierung kann beispielsweise die Überlegungen aufgegriffen werden, ob hinsichtlich demografischer Variablen wie Geschlecht, Klassenstufe oder Schulart Gruppenunterschiede in den erreichten Testwerten eine getrennte Normierung rechtfertigen. Daneben können noch weitere potentielle Merkmale bestimmt werden, für die eine nach Merkmalsausprägung getrennte Normierung notwendig sein könnte. Die Entscheidung für oder gegen solche Normierungen müssen dabei in Einklang mit der übergeordneten Zielstellung des Tests in seinem späteren Einsatzbereichs sein.

Während eine Berücksichtigung dieser mit der Testnormierung assoziierten Testgütekriterien in den bisherigen Projektschritten nicht sinnvoll gewesen wäre, so wurden die Gütekriterien der Zumutbarkeit und der Ökonomie dagegen explizit bei der Testkonstruktion berücksichtigt (vgl. Kapitel 2 und Kapitel 5). Sie sind insbesondere für die Population der Kinder mit spezifischem Förderbedarf von Relevanz und können als hinreichend umgesetzt betrachtet werden.

Daneben kommt den Testgütekriterien der Reliabilität und der Validität eine herausragende Bedeutung zu. Zur Bestimmung der Reliabilität eignen sich verschiedene Ansätze. So wurden in Kapitel 8 die inneren Konsistenzen der Testitems berichtet. Sie lagen im Bereich von .706 bis .817, was auch in Hinblick auf die inhaltliche Breite der Inhalte der Testitems für die Testreliabilität spricht. Daneben können beispielsweise noch die Test-Retest-Reliabilität und die Paralleltest-Reliabilität bestimmt werden, um den bislang entstandenen Eindruck zu untermauern. Die Paralleltest-Reliabilität erfordert jedoch mindestens zwei parallele Testformen, sie ist somit aufwändig und weniger praktikabel (Pospeschill, 2010). Die Retest-Reliabilität zielt dagegen auf die zeitliche Stabilität ab und könnte sich durch eine einfache Wiederholung der Messung umsetzen lassen.

Die Frage der Validität wurde bislang mehrfach aufgegriffen, insbesondere war sie im zentralen Fokus der in Kapitel 8 dargestellten Untersuchungen. So war die Konstruktvalidität das zentrale Thema in Kapitel 8. Dagegen spielten Überlegungen zur Inhaltsvalidität eine wichtige Rolle in der Testkonstruktion und waren die Grundlage für die Ausführungen in Kapitel 2 und Kapitel 3, in denen die zu erfassenden Konstrukte inhaltlich präzisiert wurden. Sie waren darüber hinaus bei der Entwicklung der Testmaterialen (Kapitel 5) und der Auswahl der Aufgaben für die vorläufige Endversion (Kapitel 7.4) relevant, bei der alle nach den Ergebnissen der Pilotierung sinnvollen Kombinationen aus Aufgaben- und Informationsart hinreichend berücksichtigt wurden.

Neben der Inhaltsvalidität wurde auch die Konstruktvalidität durch die Befunde aus Kapitel 8 berücksichtigt. Erste Hinweise auf eine vorhandene Konstruktvalidität können als erbracht angesehen werden. Weitere Untersuchungen zu diskriminanten und konvergenten Validitäten sind jedoch sinnvoll, um die bisherigen Ergebnisse abzusichern und zu erweitern. Die Zusammenhänge mit praktisch relevanten Kriterien bieten sich im Sinne der Kriteriumsvalidität in besonderem Maße für solche Folgeuntersuchungen an. Insbesondere sind hier Fragen zur prognostischen Validität von Interesse, die längsschnittlich umgesetzt werden kann. Ebenfalls von Bedeutung für die

Validität und insbesondere für die Konstruktvalidität ist darüber hinaus das Testformat.

Zur Validität von Tests im Multiple-Choice-Format

Die in der Validierungsuntersuchung ermittelten theoriekonformen Zusammenhänge der Lesekompetenz mit externen Lesemaßen sprechen für die Konstruktvalidität des dynamischen Lesekompetenztests und stehen damit im Widerspruch zu den Befunden von Rost und Sparfeldt (2007), die die Konstruktvalidität von Lesetests im Multiple-Choice-Format anzweifeln. Rost und Sparfeldt ließen Schüler der siebten Jahrgangsstufe Multiple-Choice-Leseverständnistestaufgaben in mehreren Versionen bearbeiten. So erhielten die Schüler neben dem Originaltest auch die Testversionen „ohne Text“, „ohne Fragen“ und „ohne Text und ohne Fragen“. Sie lösten in den Versionen „Original“, „ohne Text“ und „ohne Fragen“ mehr Aufgaben als durch reines Raten zu erwarten wäre. In der Version „ohne Text und ohne Fragen“ wurden ebenfalls teilweise mehr Aufgaben gelöst, als durch die Ratewahrscheinlichkeit zu erwarten gewesen wäre. Dies impliziert, dass die richtigen Antworten der Schüler in den Bedingungen „ohne Text“ und „ohne Text und ohne Fragen“ nicht auf die Verarbeitung des Textes und damit nicht auf Leseverstehen zurückzuführen sei. Auch können laut den Autoren Boden- und Deckeneffekte nicht als Erklärung dieses Befundmusters gelten. Die Autoren argumentieren, dass vielmehr verbale Intelligenz und vorhandenes Vorwissen für die erbrachte Testleistung hier von Relevanz sei (Rost & Sparfeldt, 2007). Diese Faktoren können aber nur in Interaktion mit dem konkret vorliegenden Material und insbesondere mit den konkreten Antwortmöglichkeiten zu überzufällig richtigen Antworten führen. Damit kommt der Rolle der Distraktoren eine besondere Bedeutung zu. Sind die Distraktoren beispielsweise so gewählt, dass sie nicht zur Fragestellung passen, dann begünstigt dies in der Bedingungen „ohne Text“ eine richtige Beantwortung der Frage. Dagegen sollten Distraktoren so gewählt werden, dass sie möglichst wenig von Wissen und Intelligenz abhängig sind, sondern insbesondere eine Verbindung zum Aufgabenstamm und zur Fragestellung aufweisen. Diesem Gedanken wurde bei der Entwicklung der Testmaterialien Rechnung getragen (vgl. Kapitel 5.1). So sollten alle Distraktoren nach

Möglichkeit im Aufgabenstamm vorkommen und ebenso wie die richtige Antwort mit der Fragestellung assoziiert werden können, beispielsweise durch eine ähnliche Funktion im Situationsmodell oder eine generelle semantische Ähnlichkeit. Die in der Validierungsuntersuchung erzielten Korrelationen der Lesekompetenz mit externen Lesemaßen sind mit Werten über .500 als hoch zu interpretieren, während die Korrelation mit dem Maß der allgemeinen kognitiven Fähigkeiten bei unter .500 lag. Damit liegt die Entscheidung für die richtige Antwort im Multiple-Choice-Test nicht nur an den allgemeinen kognitiven Fähigkeiten, sondern darüber hinaus auch an der Fähigkeit, die Informationen des Aufgabenstamms zielführend zu verarbeiten und mit den Distraktoren und der richtigen Lösung angemessen verbinden zu können. Dennoch würde sich für ein Folgeprojekt anbieten, die Distraktoren in einer weiteren Untersuchung in einer Bedingung „ohne Text“ hinsichtlich ihrer Plausibilität zu eruieren.

Insgesamt kann jedoch von angemessen guten ersten Befunden zur Validität gesprochen werden. Darüber hinaus wurden die Testgütekriterien berücksichtigt, die bis zu diesem Projektstand in sinnvoller Art und Weise hätten Berücksichtigung finden können. Damit kann von einer soliden Basis für weitere Untersuchungen ausgegangen werden. Fragen der prognostischen Validität sollten hierbei ebenso berücksichtigt werden wie die Klärung der Retest-Reliabilität und eine in Abhängigkeit vom anvisierten Einsatzfeld und der mit der Diagnostik verbundenen Zielsetzung sinnvolle Normierung. Sie sollten als nächste konkrete Schritte eingeleitet werden, um den dynamischen Lesekompetenztest weiterzuentwickeln.

9 Abschließende Würdigung und Ausblick

Das übergeordnete Ziel der vorliegenden Arbeit bestand darin, das Prinzip des dynamischen Testens mit dem Bereich der Lesekompetenz zu verbinden. Dazu wurden zunächst in einer metaanalytischen Untersuchung systematisch Befunde geprüft, welche die statisch erfasste Lesekompetenz mit der dynamischen Testung in Verbindung bringen. Ein positiver Zusammenhang zwischen dynamischem Assessment (DA) und statisch erfasster Lesekompetenz (LK) konnte nachgewiesen werden. Daneben konnte ein eigenständiges dynamisches Verfahren entwickelt und erfolgreich validiert werden, welches auch in einer Population mit spezifischem Förderbedarf vielversprechende erste Resultate in Hinblick auf die prädiktiven Korrelationen mit verschiedenen externen Kriterien lieferte.

Abschließend sei nochmals auf das Potential hingewiesen, welches dynamische Verfahren für die Beantwortung psychologischer Fragestellungen bieten. Das bislang als Forschungsgegenstand wenig berücksichtige Konstrukt der dynamischen Komponente (Sternberg & Grigorenko, 2002, S. 31) bietet trotz gewisser methodischer Herausforderungen große Potenz in der Ergänzung der kanonischen Intelligenz- und Leistungsdiagnostik. Dem Aufruf zur Verfahrensentwicklung (Wiedl, 1984) kann damit nur zugestimmt werden.

Anhang

Anhang A:	Primärstudien der metaanalytischen Untersuchung
Anhang B:	Permutationsplan der qualitativen Vorerprobung
Anhang C:	Übersicht über die Blöcke, Untergruppen und Gruppen und ihre Items
Anhang D:	Versuchsgruppen der Pilotierung: Stichprobenverteilung und Itemanzahl

Weitere Unterlagen können bei der Autorin angefordert werden.

Anhang A: Primärstudien der Metaanalyse (Pseudonym in Klammern):

Caffrey, E. (2006). *A comparison of dynamic assessment and progress monitoring in the prediction of reading achievement for students in kindergarten and first grade*. Dissertation. Vanderbilt University. (**w12**)

Compton D. L., Fuchs D., Fuchs L. S., Bouton B., Gilbert J. K., Barquero L. A. & Crouch R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*, 327–341. (**f15**)

Coventry, W. L., Byrne, B., Olson, R. K., Corley, R. & Samuelsson, S. (2011). Dynamic and static assessment of phonological awareness in preschool: a behavior-genetic study. *Journal of Learning Disabilities, 44*(4), 322–329. (**s12**)

Elleman, A. M., Compton, D. L., Fuchs, D., Fuchs, L. S. & Bouton, B. (2011). Exploring dynamic assessment as a means of identifying children at risk of developing comprehension difficulties. *Journal of Learning Disabilities, 44*, 348–357. (**s16**)

Glutting, J. J. & McDermott, P. A. (1990). Childhood learning potential as an alternative to traditional ability measures. *Psychological Assessment, 2*, 398–403. (**f27**)

Hessels, M. G. P. & Hamers, J. H. M. (1993). A learning potential test for ethnic minorities. In J. H. M. Hamers, K. Sijtsma & A. J. J. M. Ruijssenaars (Hrsg.), *Learning Potential Assessment*. Amsterdam, Swets and Zeitlinger. (**j7**)

Hippmann, K. (2008). *Prädiktoren des Schriftspracherwerbs im Deutschen*. Dissertation. Rheinisch-Westfälische Technische Hochschule Aachen. (**w2**)

Larsen, J. A. & Nippold, M. A. (2007). Morphological analysis in school-age children: Dynamic assessment of a word learning strategy. *Language, Speech, and Hearing Services in Schools, 38*(3), 201-212. (**w21**)

Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology, 84*, 353-363. (**f5**)

Speece, D. L., Cooper, D. H & Kibler, J. M. (1990). Dynamic assessment, individual differences, and academic achievement. *Learning and Individual Differences, 2*, 113–127. (**f17**)

Stanfa, K. M. (2010). *Differentiating among students: the value added of a dynamic assessment of morphological problem-solving*. Dissertation. University of Pittsburgh. (**w11**)

Swanson, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology, 84*(4), 473. (**f7**)

Swanson, H. L. (1995). Effects of dynamic testing on the classification of learning disabilities: The predictive and discriminant validity of the Swanson-Cognitive Processing Test (S-CPT). *Journal of Psychoeducational Assessment, 13*(3), 204-229. (**f6**)

Anhang B: Permutationsplan der qualitativen Vorerprobung

	Probeitems		Brückeninformationen		Lokale Informationen		Temporale Informationen		Kausale Informationen	
			4 B-Items		8 EL-Items		8 PT-Items		16 IK-Items	
Gruppe A			**Narrat. Text**	**Sachtext**	**Narrat. Text**	**Sachtext**	**Narrat. Text**	**Sachtext**	**Narrat. Text**	**Sachtext**
A1	PKNd	ITN1 ILNd	EBN1		ELN1	ELS1	PTN1	PTS1	IKN1	IKS1 IKS2 IKS9
A2	PKNd	ITN1 ILNd	EBN2		ELN2	ELS2	PTN2	PTS2	IKN2	IKS3 IKS4 IKS10
A3	PKNd	ITN1 ILNd	EBN3		ELN3	ELS3	PTN3	PTS3	IKN3	IKS5 IKS6 IKS11
A4	PKNd	ITN1 ILNd	EBN4		ELN4	ELS4	PTN4	PTS4	IKN4	IKS7 IKS8 IKS12
			6 B-Items		8 PL-Items		14 IT-Items		8 EK-Items	
Gruppe B			**Narrat. Text**	**Sachtext**	**Narrat. Text**	**Sachtext**	**Narrat. Text**	**Sachtext**	**Narrat. Text**	**Sachtext**
B1	PBN1*	ITN1 IKNd		EBS1	PLN1*	PLS1	ITN2	ITS1 ITS2 ITS3	EKN3	EKS1
B2	PBN1*	ILNd IKNd		EBS2	PLN2	PLS2	ITN1*	ITS4 ITS5 ITS6	EKN2	EKS2
B3	PBN1*	ILNd ITN1	PBN3	EBS3	PLN3	PLS3	ITN3	ITS7 ITS8	EKN1*	EKS3
B4	PTNd	ILNd IKNd	PBN2	EBS4	PLN4	PLS4	ITN4	ITS9 ITS10	EKN4	EKS4
			6 B-Items		14 IL-Items		8 ET-Items		8 PK-Items	
Gruppe C			**Narrat. Text**	**Sachtext**	**Narrat. Text**	**Sachtext**	**Narrat. Text**	**Sachtext**	**Narrat. Text**	**Sachtext**
C1	PLN1	ITN1 INKd		PBN5	ILN1	ILS1 ILS2 ILS3	ETN1	ETS1	PKN1	PKS1
C2	PLN1	ITN1 INKd		PBS2	ILN2	ILS4 ILS5 ILS6	ETN2	ETS2	PKN2	PKS2
C3	PLN1	ITN1 INKd	PBN4	PBS3	ILN3	ILS7 ILS8	ETN3	ETS3	PKN3	PKS3
C4	PLN1*	ITN1 INKd	PBN1*	PBS1	ILN4	ILS9 ILS10	ETN4	ETS4	PKN4	PKS4

Anmerkung. * *Dummy*items, die Adaption der Probeitems nötig machten und davon betroffene Probeitems

Anhang C: Übersicht über die Blöcke, Untergruppen und Gruppen und ihre Items

Gruppe	Untergruppe	Block	aufsteigende Itemreihenfolge				Untergruppe	Block	absteigende Itemreihenfolge			
A	A1	1	ELN1	IKS5	PTS1	ILS8	A2	1	ELS4	IKS6	PTS4	ILS10
		2	ELS1	PTN1	IKS1			2	ELN4	PTN4	IKS4	
		3	ELN2	PTN2	IKS2			3	ELS3	PTS3	IKN2	
		4	ELS2	IKN1	PTS2	ILS9		4	ELN3	PTN3	IKS3	
A	A2	5	ELN3	PTN3	IKS3		A1	5	ELS2	IKN1	PTS2	ILS9
		6	ELS3	PTS3	IKN2			6	ELN2	PTN2	IKS2	
		7	ELN4	PTN4	IKS4			7	ELS1	PTN1	IKS1	
		8	ELS4	IKS6	PTS4	ILS10		8	ELN1	IKS5	PTS1	ILS8
B	B1	1	EBS1	ITN1	PKN1	ITS5	B2	1	EBS4	PKN4	ITS4	
		2	EBS2	ITN2	PKN2	ITS6		2	EBN4	PKS4	ITS3	
		3	EBS3	ITN3	PKN3	ITS7		3	EBN3	PKS3	ITS2	
		4	EBN1	PKS1	ITN4			4	EBN2	PKS2	ITS1	
B	B2	5	EBN2	PKS2	ITS1		B1	5	EBN1	PKS1	ITN4	
		6	EBN3	PKS3	ITS2			6	EBS3	ITN3	PKN3	ITS7
		7	EBN4	PKS4	ITS3			7	EBS2	ITN2	PKN2	ITS6
		8	EBS4	PKN4	ITS4			8	EBS1	ITN1	PKN1	ITS5
C	C1	1	EKS1	ILN1	PBN1	ILS5	C2	1	EKS4	PBN4	ILS4	
		2	EKS2	ILN2	PBN2	ILS6		2	EKN4	PBS3	ILS3	
		3	EKS3	ILN3	PBN3	ILS7		3	EKN3	PBS4	ILS2	
		4	EKN1	PBS1	ILN4			4	EKN2	PBS2	ILS1	
C	C2	5	EKN2	PBS2	ILS1		C1	5	EKN1	PBS1	ILN4	
		6	EKN3	PBS4	ILS2			6	EKS3	ILN3	PBN3	ILS7
		7	EKN4	PBS3	ILS3			7	EKS2	ILN2	PBN2	ILS6
		8	EKS4	PBN4	ILS4			8	EKS1	ILN1	PBN1	ILS5
D	D1	1	ETS1	IKS7	PLS1	ITS8	D2	1	ETN4	IKS12	PLS4	ITS10
		2	ETN1	PLN1	IKS8			2	ETS4	PLN4	IKS11	
		3	ETN2	PLN2	IKS9			3	ETS3	PLS3	IKN4	
		4	ETS2	IKN3	PLS2	ITS9		4	ETN3	PLN3	IKS10	
D	D2	5	ETN3	PLN3	IKS10		D1	5	ETS2	IKN3	PLS2	ITS9
		6	ETS3	PLS3	IKN4			6	ETN2	PLN2	IKS9	
		7	ETS4	PLN4	IKS11			7	ETN1	PLN1	IKS8	
		8	ETN4	IKS12	PLS4	ITS10		8	ETS1	IKS7	PLS1	ITS8

Anhang D: Versuchsgruppen der Pilotierung: Stichprobenverteilung und Itemanzahl

	A1	**A2**	**B1**	**B2**	**C1**	**C2**	**D1**	**D2**
A1	-	27	29	26	29	26	28	27
A2	9	-	28	25	28	25	27	26
B1	8	7	-	27	30	27	29	28
B2	8	7	9	-	27	24	26	25
C1	10	8	8	9	-	27	29	28
C2	10	11	11	9	8	-	28	27
D1	9	12	7	8	8	7	-	27
D2	8	11	7	7	9	7	8	-

Anmerkung. Oberhalb der Diagonale ist die Anzahl der Items dargestellt, die in dieser Kombination zu bearbeiten waren. Unterhalb der Diagonale ist die jeweilige Probandenanzahl eingetragen.

Literaturverzeichnis

Achhammer, B. (2014). Pragmatische Störungen. In M. Grohnfeldt (Hrsg.), *Grundwissen der Sprachheilpädagogik und Sprachtherapie* (S. 209-214). Stuttgart: Kohlhammer.

Amelang, M., Bartussek, D., Stemmler, G. & Hagemann, D. (2006). *Differentielle Psychologie und Persönlichkeitsforschung* (Kohlhammer Standards Psychologie, 6., vollst. überarb. Aufl). Stuttgart: Kohlhammer.

Amelang, M. & Zielinski, W. (1994). *Psychologische Diagnostik und Intervention*. Berlin [u.a.]: Springer.

Artelt, C. & Dörfler, T. (2010). Förderung von Lesekompetenz als Aufgabe aller Fächer. Forschungsergebnisse und Anregungen für die Praxis. In Bayerisches Staatsministerium für Unterricht und Kultus und Staatsinstitut für Schulqualität und Bildungsforschung (Hrsg.), *ProLesen. Auf dem Weg zur Leseschule - Leseförderung in den gesellschaftswissenschaftlichen Fächern* (S. 13-36). Donauwörth: Auer Verlag.

Artelt, C., McElvany, N., Christmann, N., Richter, T. Groeben, N., Köster, J., Schneider, W., Stanat, P., Ostermeier, C , Schiefele, U. Valtin, R. & Ring, K. (2005). *Förderung von Lesekompetenz – Eine Expertise.* Bonn, Berlin: Bundesministerium für Bildung und Forschung (BMBF). Verfügbar unter http://www.bmbf.de/pub/bildungsreform_band_siebzehn.pdf

Artelt, C., Schiefele, U. & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16* (3), 363-383.

Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U. et al. (Hrsg.). (2001). *PISA 2000: Zusammenfassung zentraler Befunde*. Berlin.

Artelt, C., Drechsel, B., Bos, W. & Stubbe, T. (2008). Lesekompetenz in PISA und PIRLS/IGLU - ein Vergleich. In M. Prenzel & J. Baumert (Hrsg.), *Vertiefende Analysen zu PISA 2006. Zeitschrift für Erziehungswissenschaft Sonderheft 10/2008* (S. 35-52). Springer-Verlag.

Artelt, C., Schneider, W. & Schiefele, U. (2002). Ländervergleich zur Lesekompetenz. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M.

Prenzel, U. Schiefele et al. (Hrsg.), *Pisa 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich* (OECD, PISA, S. 55-94). Opladen: Leske + Budrich.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69-137). Opladen: Leske + Budrich.

Autenrieth, K. (2014). *Diagnostik von Textverständnis - Vorerprobung von Testitems anhand kognitiver Interviews*. Wissenschaftliche Hausarbeit. Pädagogische Hochschule Heidelberg.

Baltes, P. B. & Baltes, M. M. (1990). Psychological perspectives on successful aging: The model of selective optimization with compensation. In P. B. Baltes & M. M. Baltes (Hrsg.), *Successful aging: Perspectives from the behavioral sciences* (S. 1-34).

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A. & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61* (2), 213-238.

Baumert, J., Brunner, M., Lüdtke, O. & Trautwein, U. (2007). Was messen internationale Schulleistungsstudien? - Resultate kumulativer Wissenserwerbsprozesse. *Psychologische Rundschau, 58* (2), 118-128.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W. et al. (Hrsg.). (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Beckmann, J. F. (2001). *Zur Validierung des Konstrukts des intellektuellen Veränderungspotentials*. Berlin: logos.

Beckmann, J. F. & Guthke, J. (1999). *Psychodiagnostik des schlussfolgernden Denkens. Handbuch zur adaptiven computergestützten Intelligenz-Lerntestbatterie für Schlussfolgerndes Denken (ACIL)*. Göttingen [u.a.]: Hogrefe.

Beckmann, N. (2004). *Differentielle Effekte von Feedback in Intelligenztests*. Dissertation. Heinrich-Heine-Universität Düsseldorf.

Beckmann, N., Beckmann, J. F. & Elliott, J. G. (2009). Self-confidence and performance goal orientation interactively predict performance in a reasoning test with accuracy feedback. *Learning and Individual Differences, 19* (2), 277-282.

Belgrad, J. & Pfaff, H. (2010). Sachtexte in der Grundschule. In G. Schulz (Hrsg.), *Basisbuch Lesen* (S. 62-74). Berlin: Cornelsen-Verlag.

Bestgen, Y. & Vonk, W. (1995). The role of temporal segmentation markers in discourse processing. *Discourse Processes, 19* (3), 385-406.

Bethge, H.-J., Carlson, J. S. & Wiedl, K. H. (1982). The effects of dynamic assessment procedures on Raven matrices performance, visual search behavior, test anxiety and test orientation. *Intelligence, 6* (1), 89-97.

Blatter, K. (2014). *Familiale sprachbezogene Förderung und frühe (schrift-)sprachliche Kompetenzen. Zusammenhänge bei Kindern mit und ohne Migrationshintergrund*. Dissertation. Otto-Friedrich-Universität Bamberg.

Bliss, J. (1996). Piaget und Vygotsky: Ihre Bedeutung für das Lehren und Lernen der Naturwissenschaften. *Zeitschrift für Didaktik der Naturwissenschaften, 2* (3), 3-16.

Bodner, K. E., Engelhardt, C. R., Minshew, N. J. & Williams, D. L. (2015). Making inferences: Comprehension of physical causality, intentionality, and emotions in discourse by high-functioning older children, adolescents, and adults with autism. *Journal of autism and developmental disorders, 45* (9), 2721-2733.

Böhme, K. (2011). *Methodische und didaktische Überlegungen sowie empirische Befunde zur Erfassung sprachlicher Kompetenzen im Deutschen. Analysen zu den Bildungsstandards im Fach Deutsch für den Primarbereich*. Dissertation. Humboldt-Universität zu Berlin.

Borenstein, M., Hedges, L., Higgins, J. & Rothstein, H. (2015). Comprehensive Meta-Analysis (Version 3) [Computer software]. Verfügbar unter http://www.comprehensive.com

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler. Mit 242 Tabellen* (Springer-Lehrbuch, 6. Aufl.). Berlin: Springer.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer Medizin Verlag.

Bortz, J., Lienert, G. A. & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik*. Heidelberg: Springer Medizin Verlag Heidelberg.

Bos, W., Valtin, R., Lankes, E.-M., Schwippert, K., Voss, A., Badel, I. et al. (2004). Lesekompetenzen am Ende der vierten Jahrgangsstufe in einigen Ländern der Bundesrepublik Deutschland im nationalen und internationalen Vergleich. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin & G. Walther (Hrsg.), *IGLU. Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (S. 49-92). Münster [u.a.]: Waxmann.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G. & Valtin, R. (Hrsg.). (2003). *Erste Ergebnisse aus IGLU : Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Lankes, E.-M., Schwippert, K., Valtin, R., Voss, A., Badel, I. et al. (2003). Lesekompetenzen deutscher Grundschülerinnen und Grundschüler am Ende der vierten Jahrgangsstufe im internationalen Vergleich. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, G. Walther & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU : Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 69-142). Münster: Waxmann.

Brandes, H. (2013). Entwicklung, soziokultureller Ansatz nach Wygotski. In M. A. Wirtz & J. Strohmer (Hrsg.), *Dorsch. Lexikon der Psychologie* (16. vollst. überarb. Aufl., S. 467). Bern: H. Huber.

Brown, A. L., Campione, J. C., Webber, L. S. & McGilly, K. (1992). Interactive learning environments: A new look at assessment and instruction. In B. R. Gifford & M. C. O'Connor (Hrsg.), *Changing assessments. Alternative views of aptitude, achievement and instruction* (S. 121-211). New York: Springer Science+Business Media, LLC.

Brown, A. L. & Ferrara, R. A. (1999). Diagnosing zones of proximal development. In P. Lloyd & C. Fernyhough (Hrsg.), *Lev Vygotsky critical assessments. Volume III The zone of proximal development* (S. 225-253). London: Routledge.

Brown, H. M., Oram-Cardy, J. & Johnson, A. (2013). A meta-analysis of the reading comprehension skills of individuals on the autism spectrum. *Journal of autism and developmental disorders, 43* (4), 932-955.

Budoff, M. & Corman, L. (1976). Effectiveness of a learning potential procedure in improving problem-solving skills of retarded and nonretarded children. *Journal of Mental Deficiency, 81* (3), 260-264.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München [u.a.]: Pearson.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience, 14* (5), 365-376.

Caffrey, E. (2006). *A comparison of dynamic assessment and progress monitoring in the predicition of reading achievement for students in kindergarten and first grade*. Dissertation. Vanderbilt University.

Caffrey, E., Fuchs, D. & Fuchs, L. S. (2008). The predictive validity of dynamic assessment: A review. *The Journal of Special Education, 41* (4), 254-270.

Caillies, S., Denhière, G. & Kintsch, W. (2002). The effect of prior knowledge on understanding from text: Evidence from primed recognition. *European Journal of Cognitive Psychology, 14* (2), 267-286.

Caledo, M. D. & Márquez, J. (1998). Psychometric properties of a learning potential test for reading: The Picture Word Game. *European Journal of Psychological Assessment, 14,* 124-133.

Calvo, M. G. & Carreiras, M. (1993). Selective influence of test anxiety on reading processes. *British Journal of Psychology, 84,* 375-388.

Campione, J. C. & Brown, A. L. (1985). *Dynamic assessment: One approach and some initial data.* Technical Report No. 361.

Carlson, J. S. & Wiedl, K. H. (1979). Toward a differential testing approach: Testing-the-limits employing the Raven matrices. *Intelligence, 3* (4), 323-344.

Carlson, J. S. & Wiedl, K. H. (1992). Principles of dynamic assessment: The application of a specific model. *Learning and Individual Differences, 4* (2), 153-166.

Cassady, J. C. & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27* (2), 270-295.

Chamorro-Premuzic, T. & Furnham, A. (2003). Personality traits and academic examination performance. *European Journal of Personality, 17* (3), 237-250.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A. et al. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97* (2), 268-274.

Chikalanga, I. (1992). A suggested taxonomy of inferences for the reading teacher. *Reading in a Foreign Language, 8* (2), 697-709.

Christmann, U. & Groeben, N. (1999). Psychologie des Lesens. In B. Franzmann, K. Hasemann, D. Löffler & E. Schön (Hrsg.), *Handbuch Lesen* (Genehmigte Lizenzausg., S. 145-223). München: Saur.

Christmann, U. & Groeben, N. (2006). Anforderungen und Einflussfaktoren bei Sach- und Informationstexten. In N. Groeben & B. Hurrelmann (Hrsg.),

Lesekompetenz. Bedingungen, Dimensionen, Funktionen (2. Aufl., S. 150-173). Weinheim und München: Juventa Verlag.

Christmann, U. (2006). Textverstehen. In J. Funke & P. A. Frensch (Hrsg.), *Handbuch der Allgemeinen Psychologie - Kognition* (Handbuch der Psychologie, Bd. 5, S. 612-620). Göttingen: Hogrefe.

Coiro, J. (2011). Predicting reading comprehension on the internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research: A publication of the Literacy Research Association, 43* (4), 352-392.

Compton, D. L. (2006). How should "unresponsiveness" to secondary intervention be operationalized? It is all about the nudge. *Journal of Learning Disabilities, 39* (2), 170-173.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A. et al. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102,* 327-341.

Conrad, F., Blair, J. & Tracy, E. (1999). Verbal reports are data! A theoretical approach to cognitive interviews. *Proceedings of the Federal committee on Statistical Methodology Research Conference,* 11-20.

Cooper, H. M. (2010). *Research synthesis and meta-analysis. A step-by-step approach* (Applied social research methods series, Bd. 2, 4. Aufl.). Los Angeles: Sage.

Coté, N. & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. In H. v. Oostendorp & S. R. Goldman (Hrsg.), *The construction of mental representations during reading* (S. 169-193). Mahwah: L. Erlbaum Associates.

Coventry, W. L., Byrne, B., Olson, R. K., Corley, R. & Samuelsson, S. (2011). Dynamic and static assessment of phonological awareness in preschool: a behavior-genetic study. *Journal of Learning Disabilities, 44* (4), 322-329.

Darhower, M. A. (2014). Synchronous computer-mediated dynamic assessment. A case study of L2 Spanish past narration. *CALICO Journal, 31* (2), 221-243.

De Jonge, P. & De Jong, P. F. (1996). Working memory, intelligence and reading ability in children. *Personality and Individual Differences, 21* (6), 1007-1020.

Dopkins, S. (1996). Representation of superordinate goal inferences in memory. *Discourse Processes, 21* (1), 85-104.

Dörfler, T., Golke, S. & Artelt, C. (2009). Dynamic assessment and its potential for the assessment of reading competence. *Studies in Educational Evaluation, 35,* 77-82.

Dörfler, T., Golke, S. & Artelt, C. (in press). Evaluating prerequisites for the development of a dynamic test of reading competence: Feedback effects on reading comprehension in children. In D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (Hrsg.), *Competence Assessment in Education: Research, Models and Instruments.* Heidelberg: Springer.

Dörfler, T., Golke, S. & Artelt, C. (2010). Dynamisches Testen der Lesekompetenz. Konzeption und erste Ergebnisse. *Zeitschrift für Pädagogik, 56*(2), 154-164.

Drechsel, B. & Artelt, C. (2007). Lesekompetenz. In M. Prenzel (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 225-248). Münster, München [u.a.]: Waxmann.

Eckert, C. (2012). *Beeinflusst Stereotype Threat die Leseleistung von Jungen?* Dissertation. Johannes Gutenberg-Universität Mainz.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z. & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36* (3), 250-287.

Eid, M., Gollwitzer, M. & Schmitt, M. (2011). *Statistik und Forschungsmethoden* (2. Aufl.). Weinheim, Basel: Beltz.

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen [u.a.]: Hogrefe.

Einhaus, E. & Schecker, H. (2007). Schülerleistungen und Kompetenzmodellierung. In D. Höttecke (Hrsg.), *Naturwissenschaftlicher Unterricht im internationalen Vergleich* (S. 154-157). Berlin: Lit Verlag.

Elbro, C. (1996). Early linguistic abilities and reading development: A review and a hypothesis. *Reading and Writing: An Interdisciplinary Journal, 8,* 453-485.

Elleman, A. M., Compton, D. L., Fuchs, D., Fuchs, L. S. & Bouton, B. (2011). Exploring dynamic assessment as a means of identifying children at risk of developing comprehension difficulties. *Journal of Learning Disabilities, 44,* 348-357.

Ellert, C. (2015). *Differentialpsychologische Analysen zur Lesekompetenz in der Grundschule*. Wissenschaftliche Hausarbeit. Pädagogische Hochschule Heidelberg.

Elley, W. B. (1994). *The IEA Study of reading literacy: Achievement and instruction in thirty-two school systems.* Oxford: Pergamon Press.

Elliott, J. (2003). Dynamic assessment in educational settings: Realising potential. *Educational Review, 55* (1), 15-32.

Eysenck, M. W., Derakshan, N., Santos, R. & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion, 7* (2), 336-353.

Eysenck, M. W. & Keane, M. T. (2006). *Cognitive psychology: a student's handbook* (5. Aufl.). Hove [u.a.]: Psychology Press.

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (2011). *Statistik. Der Weg zur Datenanalyse* (Springer-Lehrbuch, 7. Aufl.). Berlin [u.a.]: Springer.

Faul, F., Buchner, A., Erdfelder, E. & Lang, A. G. (2008). G*Power (Version 3.0.10) [Computer software]. Verfügbar unter http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

Feng, X. & Rost, D. H. (2015). Selbstberichtete Zeugnisdaten: Weitere Evidenz für ihre (partielle) Brauchbarkeit. *Psychologie in Erziehung und Unterricht, 62* (4), 253-264.

Ferrara, R. A., Brown, A. L. & Campione, J. C. (1986). Children's learning and transfer of inductive reasoning rules: Studies of proximal development. *Child Development, 57,* 1087-1099.

Fischer, C. (2009). *Texte, Gattungen, Textsorten und ihre Verwendung in Lesebüchern*. Dissertation. Justus-Liebig-Universität Gießen.

Fischer, M. Y. & Pfost, M. (2015). Wie effektiv sind Maßnahmen zur Förderung der phonologischen Bewusstheit? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47* (1), 35-51.

Fisseni, H.-J. (2004). *Lehrbuch der psychologischen Diagnostik*. Göttingen [u.a.]: Hogrefe.

Fuchs, D., Compton, D. L., Fuchs, L. S., Bouton, B. & Caffrey, E. (2011). The construct and predictive validity of a dynamic assessment of young children learning to read: implications for RTI frameworks. *Journal of Learning Disabilities, 44* (4), 339-347.

Giel, B. (2014). Sprachentwicklungsstörungen im Zusammenhang mit anderen Entwicklungsbedingungen. In M. Grohnfeldt (Hrsg.), *Grundwissen der Sprachheilpädagogik und Sprachtherapie* (S. 215-219). Stuttgart: Kohlhammer.

Glück, C. W. (2003). Semantisch-lexikalische Störung als Teilsymptomatik von Sprachentwicklungsstörungen. In M. Grohnfeldt (Hrsg.), *Lehrbuch der Sprachheilpädagogik und Logopädie: Band 2* (2. Aufl., S. 75-87). Stuttgart: Kohlhammer.

Glück, C. W. & Elsing, C. (2014). Semantisch-lexikalische Störungen. In M. Grohnfeldt (Hrsg.), *Grundwissen der Sprachheilpädagogik und Sprachtherapie* (S. 204-208). Stuttgart: Kohlhammer.

Glutting, J. J. & McDermott, P. A. (1990). Childhood learning potential as an alternative to traditional ability measures. *Psychological Assessment, 2,* 398-403.

Goldhammer, F. & Hartig, J. (2012). Interpretation von Testresultaten und Testeichung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 173-202). Berlin [u.a.]: Springer.

Golke, S., Matthäi, J. & Artelt, C. (2013). Sachtexte lesen und verstehen: Textverstehen aus psychologischer Perspektive. *Der Deutschunterricht: Beiträge zu seiner Praxis und wissenschaftlichen Grundlegung, 65* (6), 19-29.

Golke, S. (2013). *Effekte elaborierter Feedbacks auf das Textverstehen. Untersuchungen zur Wirksamkeit von Feedbackinhalten unter Berücksichtigung des Präsentationsmodus in computerbasierten Testsettings*. Dissertation. Otto-Friedrich-Universität Bamberg.

Golke, S., Dörfler, T. & Artelt, C. (2015). The impact of elaborated feedback on text comprehension within a computer-based assessment. *Learning and Instruction, 39,* 123-136.

Göpferich, S. (2007). *Praktische Handreichung für Studien mit lautem Denken und Translog (2000 und 2006),* Institut für Theoretische und Angewandte Translationswissenschaft (ITAT), Karl-Franzens-Universität Graz. Verfügbar unter http://www.susanne-goepferich.de/Handreichung.pdf

Graesser, A. C., Millis, K. K. & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48,* 163-189.

Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101* (3), 371-395.

Grigorenko, E. L. (2004). Is it possible to study intelligence wthout using the concept of intelligence? An example from Soviet/Russian psychology. In R. J. Sternberg (Hrsg.), *International handbook of intelligence* (S. 170-211). Cambridge, UK, New York, NY [u.a.]: Cambridge University Press.

Grigorenko, E. L. (2009). Dynamic assessment and response to intervention: two sides of one coin. *Journal of Learning Disabilities, 42* (2), 111-132.

Grigorenko, E. L. & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124* (1), 75-111.

Grigorenko, E. L., Sternberg, R. J., Jukes, M., Alcock, K., Lambo, J., Ngorosho, D. et al. (2006). Effects of antiparasitic treatment on dynamically and statically tested cognitive skills over time. *Journal of Applied Developmental Psychology, 27* (6), 499-526.

Grills-Taquechel, A. E., Fletcher, J. M., Vaughn, S. R. & Stuebing, K. K. (2012). Anxiety and reading difficulties in early elementary school: evidence for unidirectional- or bi-directional relations? *Child Psychiatry & Human Development, 43* (1), 35-47.

Groenwald, E. (2012). *Empirische Untersuchung der Interessen von Mädchen und Jungen im Grundschulalter zu Inhalten des naturwissenschaftlichen Sachunterrichts durch altersangemessene Fragebögen und qualitative Interviews*. Masterarbeit. Carl von Ossietzky Universität Oldenburg.

Grohnfeldt, M. & Ritterfeld, U. (2003). Grundlagen der Sprachheilpädagogik und Logopädie. In M. Grohnfeldt (Hrsg.), *Lehrbuch der Sprachheilpädagogik und Logopädie: Band 2* (2. Aufl., S. 15-45). Stuttgart: Kohlhammer.

Gustafson, S., Svensson, I. & Fälth, L. (2014). Response to intervention and dynamic assessment: Implementing systematic, dynamic and individualised interventions in primary school. *International Journal of Disability, Development and Education, 61* (1), 27-43.

Guthke, J. & Gitter, K. (1991). Prognose der Schulleistungsentwicklung mittels Status- und Lerntests in der Vorschulzeit. In H. Teichmann, B. Meyer-Probst & D. Roether (Hrsg.), *Risikobewältigung in der lebenslangen psychischen Entwicklung* (S. 141-147). Berlin: Verlag Gesundheit.

Guthke, J., Klauer, C. & Vahle, H. (2002). Prüfung der inneren Validität bei adaptiven Kurzzeit-Lerntests durch probabilistische Testmodelle. *Zeitschrift für Differentielle und Diagnostische Psychologie, 23* (2), 113-127.

Guthke, J. & Stein, H. (1996). Are learning tests the better version of intelligence tests? *European Journal of Psychological Assessment, 12* (1), 1-13.

Guthke, J., Wolschke, P., Willmes, K. & Huber, W. (1992). Leipziger Lerntest - Diagnostisches Programm zum begriffsanalogen Klassifizieren (DP-BAK). *Heilpädagogische Forschung, 18,* 153-161.

Guthke, J. (1992). Learning tests-the concept, main research findings, problems and trends. *Learning and Individual Differences, 4* (2), 137-151.

Guthke, J., Beckmann, J. F. & Wiedl, K. H. (2003). Dynamik im dynamischen Testen. *Psychologische Rundschau, 54* (4), 225-232.

Guthke, J. & Wiedl, K. H. (1996). *Dynamisches Testen: zur Psychodiagnostik der intraindividuellen Variabilität; Grundlagen, Verfahren und Anwendungsfelder*. Göttingen [u.a.]: Hogrefe, Verl. für Psychologie.

Häcker, H. (2013). Dynamische Testdiagnostik. In M. A. Wirtz & J. Strohmer (Hrsg.), *Dorsch. Lexikon der Psychologie* (16. vollst. überarb. Aufl., S. 1540). Bern: H. Huber.

Hakala, C. M. & O'Brien, E. J. (1995). Strategies for resolving coherence breaks in reading. *Discourse Processes, 20* (2), 167-185.

Hall, C. S. (2015). Inference instruction for struggling readers: a synthesis of intervention research. *Educational Psychology Review,* 1-22.

Hänsel, D. (2003). Die Sonderschule - ein blinder Fleck in der Schulsystemforschung. *Zeitschrift für Pädagogik, 49,* 591-609.

Harley, T. A. (2008). *The psychology of language. From data to theory*. New York: Psychology Press.

Harris, S. (2014). *The effects of a test-taking skills intervention on test anxiety and test performance of 4th graders*. Masterarbeit. Louisiana State University.

Hassaskhah, J. & Haghparast, M. J. (2012). A comparative study of the impact of DA models on the writing ability and attitude of Iranian EFL learners. *The Buckingham Journal of Language and Linguistics, 5,* 38-51.

Hatcher, P. J. (2000). Predictors of reading recovery book levels. *Journal of Research in Reading, 23* (1), 67-77.

Hatz, H. (2015). *Phonologische Bewusstheit und Schriftspracherwerb. Auswirkungen eines Trainings phonologischer Bewusstheit und eines um Rechtschreibinhalte erweiterten Trainings im ersten Schuljahr auf den Erwerb des Lesens und Rechtschreibens bei Schülerinnen und Schülern mit gering ausgebildeten schriftsprachspezifischen Vorläuferfertigkeiten*. Dissertation. Pädagogische Hochschule Heidelberg.

Hedges, L. V. & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3* (4), 486-504.

Heinen, S. (2001). *Der Einfluss von Vorwissen, Interesse und Arbeitsgedächtniskapazität auf die mentale Repräsentation von Texten*. Dissertation. Universität Bielefeld.

Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+ R).* Beltz test.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58* (1), 47-77.

Hessels, M. G. P. (1997). Low IQ but high learning potential: why Zeyneb and Moussa do not belong in special education. *Educational and Child Psychology, 14* (4), 121-136.

Hessels, M. G. P. & Hamers, J. H. M. (1993). A learning potential test for ethnic minorities. In J. H. M. Hamers, K. Sijtsma & A. J. J. M. Ruijssenaars (Hrsg.), *Learning potential assessment.* Amsterdam: Swets and Zeitlinger.

Hiebert, E. H. (1999). Text matters in learning to read. *The Reading Teacher, 52* (6), 552-566.

Higgins, J. P. T. & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539-1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.), 327,* 557-560.

Hippmann, K. (2008). *Prädiktoren des Schriftspracherwerbs im Deutschen*. Dissertation. Rheinisch-Westfälische Technische Hochschule Aachen.

Hohm, M. (2005). *Zum Zusammenhang von Sprachbewusstheit, Lesekompetenz und Textverstehen*. Dissertation. Julius-Maximilians-Universität Würzburg.

Hohn, K., Schiepe-Tiska, A., Sälzer, C. & Artelt, C. (2013). Lesekompetenz in PISA 2012: Veränderungen und Perspektiven. In M. Prenzel, C. Sälzer, E. Klieme & O. Köller (Hrsg.), *PISA 2012. Fortschritte und Herausforderungen in Deutschland* (S. 217-244). Münster [u.a.]: Waxmann.

Hossiep, R., Turck, D. & Hasella, M. (2002). Bochumer Matrizentest (BOMAT - advanced). In E. Brähler, H. Holling, D. Leutner & F. Petermann (Hrsg.), *Brickenkamp Handbuch psychologischer und pädagogischer Tests. Band 1* (S. 115-116). Göttingen: Hogrefe.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F. & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological Methods, 11* (2), 193-206.

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis. Correcting error and bias in research findings* (2nd ed). Thousand Oaks, Calif: Sage.

Hurley, E. & Murphy, R. (2015). The development of a new method of idiographic measurement for dynamic assessment intervention. *Journal of Pedagogy, 6* (1), 43-60.

Hurrelmann, B. (2002). Prototypische Merkmale der Lesekompetenz. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz. Bedingungen, Dimensionen, Funktionen* (S. 275-286). Weinheim und München: Juventa Verlag.

Hurrelmann, B. (2006). Prototypische Merkmale der Lesekompetenz. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz. Bedingungen, Dimensionen, Funktionen* (2. Aufl., S. 275-286). Weinheim und München: Juventa Verlag.

IBM Corporation. (2012). IBM SPSS Statistics for Windows (Version 21.0) [Computer software]. Armonk, NY: IBM Corp.

Jacobs, E. L. (2001). The effects of adding dynamic assessment components to a computerized preschool language screening test. *Communication Disorders Quarterly, 22* (4), 217-226.

Jahani, A. (2014). The effects of text length in terms of redundancy on reading comprehension of Iranian L2 learners. *Молодой ученый, 65*, 590-593.

Javanbakht, N. & Hadian, M. (2014). The effects of test anxiety on learners' reading test performance. *Procedia - Social and Behavioral Sciences, 98,* 775-783.

Jeltova, I., Birney, D., Fredine, N., Jarvin, L., Sternberg, R. J. & Grigorenko, E. L. (2011). Making instruction and assessment responsive to diverse students' progress: group-administered dynamic assessment in teaching mathematics. *Journal of Learning Disabilities, 44* (4), 381-395.

Jensen, A. R. (1989). The relationship between learning and intelligence. *Learning and Individual Differences, 1,* 37-62.

Johnson, W. L. & Rickel, J. W. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *Journal of Artificial Intelligence in Education*, 47-78.

Johnson, W., Bouchard, T. J., Segal, N. L. & Samuels, J. (2005). General intelligence and reading performance in adults: is the genetic factor structure the same as for children? *Personality and Individual Differences, 38* (6), 1413-1428.

Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 27-74). Berlin [u.a.]: Springer.

Kannengieser, S. (2014). Spezifische Sprachentwicklungsstörungen. In M. Grohnfeldt (Hrsg.), *Grundwissen der Sprachheilpädagogik und Sprachtherapie* (S. 188-198). Stuttgart: Kohlhammer.

Kantor, P. T., Wagner, R. K., Torgesen, J. K. & Rashotte, C. A. (2011). Comparing two forms of dynamic assessment and traditional assessment of preschool phonological awareness. *Journal of Learning Disabilities, 44* (4), 313-321.

Karing, C., Pfost, M. & Artelt, C. (2013). Is secondary school teacher judgment accuracy related to the development of students' reading literacy? In M. Pfost, C. Artelt & S. Weinert (Hrsg.), *The development of reading literacy from early childhood to adolescence. Empirical findings from the Bamberg BiKS longitudinal studies* (Schriften aus der Fakultät Humanwissenschaften der Otto-Friedrich-Universität Bamberg, Bd. 14, S. 279-310). Bamberg: Univ. of Bamberg Press.

Kendeou, P. & van den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition, 35* (7), 1567-1577.

Kessler, H. (2015). *Kurzlehrbuch Medizinische Psychologie und Soziologie*. Stuttgart, New York: Thieme.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review, 95* (2), 163-182.

Kintsch, W. (1992). How readers construct situation models for stories: The role of syntactic cues and causal inferences. In A. F. Healy, S. M. Kosslyn & R. M. Shriffrin (Hrsg.), *From learning processes to cognitive processes. Essays in honor of William K. Estes* (S. 261-278). Hillsdale, NJ: Erlbaum.

Kintsch, W. (1998). *Comprehension. A paradigm for cognition*. Cambridge, New York: Cambridge University Press.

Kirsch, I. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured*. Princton, NJ: Educational Testing Service.

Klauer, K. C. & Sydow, H. (1992). Interindividuelle Unterschiede in der Lernfähigkeit. Zur Analyse von Lernprozessen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 3*, 175-190.

Klicpera, C. & Gasteiger-Klicpera, B. (1998). *Psychologie der Lese- und Schreibschwierigkeiten. Entwicklung, Ursachen, Förderung* (2. Aufl.). Weinheim: Beltz.

Klicpera, C., Schabmann, A. & Gasteiger-Klicpera, B. (2010). *Legasthenie - LRS. Modelle, Diagnose, Therapie und Förderung; mit 100 Übungsfragen* (3., aktualisierte Aufl.). München, Basel: Reinhardt.

Klime, E. (2004). Was sind Kompetenzen und wie lassen sie sich messen? *Auszug aus Pädagogik, 6* (10-13).

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119* (2), 254-284.

Kovalčíková, I. (2015). From dynamic assessment of cognitive abilities to educational interventions: Trends in cognitive education. *Journal of Pedagogy, 6* (1), 5-21.

Krach, K. S., McCreery, M. P., Loe, S. A. & Jones, W. P. (2015). Do dispositional characteristics influence reading? Examining the impact of personality on reading fluency. *Reading Psychology,* 1-17.

Krohne, H. W. & Hock, M. (2007). *Psychologische Diagnostik. Grundlagen und Anwendungsfelder* (Kohlhammer Standards Psychologie, 1. Aufl). Stuttgart: Kohlhammer.

Kromrey, J. D. & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: an empirical comparison of common missing-data treatments. *Educational and Psychological Measurement, 54* (3).

Krüsken, J. (2007). Entwicklung von Schülerleistungen und Zensuren in der Grundschule. In H. Ditton (Hrsg.), *Kompetenzaufbau und Laufbahnen im Schulsystem. Ergebnisse einer Längsschnittuntersuchung an Grundschulen* (S. 41-62). Münster: Waxmann.

Küspert, P. & Schneider, W. (2002). Würzburger Leise Leseprobe (WLLP). In E. Brähler, H. Holling, D. Leutner & F. Petermann (Hrsg.), *Brickenkamp Handbuch psychologischer und pädagogischer Tests. Band 1* (S. 393-394). Göttingen: Hogrefe.

Larsen, J. A. & Nippold, M. A. (2007). Morphological analysis in school-age children: Dynamic assessment of a word learning strategy. *Language, Speech, and Hearing Services in Schools, 38* (3), 201-212.

Lawrence, N. & Cahill, S. (2014). The impact of dynamic assessment: an exploration of the views of children, parents and teachers. *British Journal of Special Education, 41* (2), 191-211.

Lehmann, R. H. & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin:* Berlin: Senatsverwaltung für Bildung, Jugend und Sport.

Lehmann, R. H., Peek, R. & Poerschke, J. (2006). *Hamburger Lesetest für 3. und 4. Klassen (HAMLET 3-4)* (2. Aufl.). Göttingen [u.a.]: Hogrefe.

Lehmann, R. H., Peek, R. & von Stritzky, R. (1995). *Leseverständnis und Lesegewohnheiten deutscher Schülerinnen und Schüler*. Weinheim: Beltz.

Lenhard, W. (2013). *Leseverständnis und Lesekompetenz. Grundlagen - Diagnostik - Förderung* (Lehren und Lernen). Stuttgart: Kohlhammer.

Lenhard, W. & Artelt, C. (2009). Komponenten des Leseverständnisses. In W. Lenhard & W. Schneider (Hrsg.), *Diagnostik und Förderung des Leseverständnisses* (Tests und Trends, Bd. 7, S. 1-17). Göttingen [u.a.]: Hogrefe.

Lenhard, W. & Lenhard, A. (2014). *Signifikanztests bei Korrelationen*. Verfügbar unter www.psychometrica.de/korrelation.html. Bibergau: Psychometrica.

Lindner, M. A., Strobel, B. & Köller, O. (2015). Multiple-Choice-Prüfungen an Hochschulen? *Zeitschrift für Pädagogische Psychologie, 29* (3-4), 133-149.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis* (Applied social research methods series, v. 49). Thousand Oaks, Calif: Sage Publications.

Lohaus, A. & Vierhaus, M. (2015). *Entwicklungspsychologie des Kindes- und Jugendalters für Bachelor*. Berlin [u.a.]: Springer.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. *Psychologische Rundschau, 58,* 103-117.

Mammarella, I. C., Ghisi, M., Bomba, M., Bottesi, G., Caviola, S., Broggi, F. et al. (2016). Anxiety and depression in children with nonverbal learning disabilities, reading disabilities, or typical development. *Journal of Learning Disabilities, 49* (2), 130-139.

Mardani, M. & Tavakoli, M. (2011). Beyond reading comprehension: The effect of adding a dynamic assessment component on EFL reading comprehension. *Journal of Language Teaching and Research, 2* (3), 688-696.

Marx, H. (2002). Knuspels Leseaufgaben (KNUSPEL-L). In E. Brähler, H. Holling, D. Leutner & F. Petermann (Hrsg.), *Brickenkamp Handbuch psychologischer und pädagogischer Tests. Band 1* (S. 383-385). Göttingen: Hogrefe.

Mason, R. A. & Just, M. A. (2004). How the brain processes causal inferences in text. *Psychological Science, 15* (1-7).

Mathieu, M. (2014). *Aufgabenbezogene Leistung in ERP-gestützten Arbeitsprozessen: Eine empirische Analyse am Beispiel der dispositiven Auftragsbearbeitung*. Wiesbaden: Springer Gabler.

Mayringer, H. & Wimmer, H. (2005). *Salzburger Lese-Screening für die Klassenstufen 1-4. (SLS 1-4)*. Bern: Hans Huber.

McElvany, N. & Schneider, C. (2009). Förderung von Lesekompetenz. In W. Lenhard & W. Schneider (Hrsg.), *Diagnostik und Förderung des Leseverständnisses* (Tests und Trends, Bd. 7, S. 151-183). Göttingen [u.a.]: Hogrefe.

McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99* (3), 440-466.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55,* 51-62.

McNamara, T. P., Miller, D. L. & Bransford, J. D. (1996). Mental models and reading comprehension. In R. Barr, M. L. Kamil, P. B. Mosenthal & P. D.

Pearson (Hrsg.), *Handbook of reading research* (S. 490-511). Mahwah, NJ: Erlbaum Assoc.

Mehri, E. & Amerian, M. (2015). Challenges to dynamic assessment in second language learning. *Theory and Practice in Language Studies, 5* (7), 1458-1466.

Meijer, J. (2001). Learning potential and anxious tendency: Test anxiety as a bias factor in educational testing. *Anxiety, Stress & Coping: An International Journal, 14* (3), 337-362.

Melby-Lervåg, M., Lyster, S.-A. H. & Hulme, C. (2012). Phonological skills and their role in learning to read: a meta-analytic review. *Psychological Bulletin, 138* (2), 322-352.

Mesmer, H. A., Cunningham, J. W. & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly, 47* (3), 235-258.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7-26). Berlin [u.a.]: Springer.

Moutafi, J., Furnham, A. & Tsaousis, I. (2006). Is the relationship between intelligence and trait Neuroticism mediated by test anxiety? *Personality and Individual Differences, 40* (3), 587-597.

Müller, B. & Richter, T. (2013). Lesekompetenz. In J. Gabrowski (Hrsg.), *Sinn und Unsinn von Kompetenzen: Fähigkeitskonzepte im Bereich von Sprache, Medien und Kultur* (S. 29-49). Leverkusen: Budrich.

Murphy, R. & Maree, D. J. F. (2006). A review of South African research in the field of dynamic assessment. *South African Journal of Psychology, 36* (1), 168-191.

Murphy, R. (2011). *Dynamic assessment, intelligence and measurement*. Chichester et al.: John Wiley & Sons.

Narciss, S. (2006). *Informatives tutorielles Feedback. Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 56). Münster: Waxmann.

Nold, G. & Willenberg, H. (2007). Lesefähigkeit. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Konsequenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 23-41). Weinheim [u.a.]: Beltz.

Noordman, L. G. M. & Vonk, W. (1998). Memory-based processing in understanding causal information. *Discourse Processes, 26* (2-3), 191-212.

Oakhill, J. & Cain, K. (2004). The development of comprehension skills. In T. Nunes & P. Bryant (Hrsg.), *Handbook of Children's Literacy* (S. 155-180). Dordrecht: Kluwer Academic Publishers.

OECD & Statistics Canada. (1995). *Literacy, economy and society. Results of the first International Adult Literacy Survey.* Paris and Ottawa: OECD and Statistics Canada.

Osburg, C. (2003). Sprachentwicklungsstörungen und Störungen des Schriftspracherwerbs. In M. Grohnfeldt (Hrsg.), *Lehrbuch der Sprachheilpädagogik und Logopädie: Band 2* (2. Aufl., S. 113-125). Stuttgart: Kohlhammer.

Ozuru, Y., Dempsey, K. & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19* (3), 228-242.

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences, 13* (3), 423-451.

Pearson, P. D. & Hiebert, E. H. (2014). The state of the field. *The Elementary School Journal, 115* (2), 161-183.

Peltenburg, M., van den Heuvel-Panhuizen, M. & Doig, B. (2009). Mathematical power of special-needs pupils: An ICT-based dynamic

assessment format to reveal weak pupils' learning potential. *British Journal of Educational Technology, 40* (2), 273-284.

Pena, E., Iglesias, A. & Lidz, C. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10* (2), 138-154.

Peterson, R. A. & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *The Journal of Applied Psychology, 90* (1), 175-181.

Pfost, M. (2015). Children's phonological awareness as a predictor of reading and spelling. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47* (3), 123-138.

Pishghadam, R., Barabadi, E. & Kamrood, A. M. (2011). The differing effect of computerized dynamic assessment of L2 reading comprehension on high and low achievers. *Journal of Language Teaching and Research, 2* (6), 1353-1358.

Poehner, M. E. (2008). *Dynamic assessment. A Vygotskian approach to understanding and promoting L2 development* (Bd. 9): Springer.

Pospeschill, M. (2009). *SPSS für Fortgeschrittene. Durchführung fortgeschrittener statistischer Analysen:* RRZN.

Pospeschill, M. (2010). *Testtheorie, Testkonstruktion, Testevaluation. Mit 77 Fragen zur Wiederholung*. München [u.a.]: Reinhardt.

Pospeschill, M. & Spinath, F. M. (2009). *Psychologische Diagnostik*. München: Reinhardt.

Preckel, F. (2002). Advanced Progressive Matrices (APM). In E. Brähler, H. Holling, D. Leutner & F. Petermann (Hrsg.), *Brickenkamp Handbuch psychologischer und pädagogischer Tests. Band 1* (S. 93-96). Göttingen: Hogrefe.

Prenzel, M., Carstensen, C. H., Frey, A., Drechsel, B. & Rönnebeck, S. (2007). PISA 2006 - Eine Einführung in die Studie. In M. Prenzel (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 31-60). Münster, München [u.a.]: Waxmann.

Priebe, S. J., Keenan, J. M. & Miller, A. C. (2012). How prior knowledge affects word identification and comprehension. *Reading and Writing, 25* (1), 131-149.

Prüfer, P. & Rexroth, M. (2005). Kognitive Interviews. *ZUMA How-to-Reihe, 15* (121), 95-116.

Pruisken, C. (2005). Grundschüler und ihre Freizeit: Sind Kinder heute gering und einseitig interessiert? *Unterrichtswissenschaft, 33* (3), 272-288.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software]. Wien, Österreich: R Foundation for Statistical Computing. Verfügbar unter http://www.R-project.org

Rack, O. & Christophersen, T. (2009). Experimente. In S. Albers, D. Klapper, U. Konradt, A. Walter & J. Wolf (Hrsg.), *Methodik der empirischen Forschung* (S. 17-32). Wiesbaden: Springer Fachmedien.

Ramazanpour, G., Nourdad, N. & Nouri, N. (2016). Gender differences in the effect of dynamic assessment on grammatical accuracy of writings. *Theory and Practice in Language Studies, 6* (1), 90-96.

Rapp, A. F. (2013). Zone der nächsten Entwicklung. In M. A. Wirtz & J. Strohmer (Hrsg.), *Dorsch. Lexikon der Psychologie* (16. vollst. überarb. Aufl., S. 1705-1706). Bern: H. Huber.

Rasinski, T. & Padak, N. (2015). *Research to practice: Text considerations in literacy teaching and learning,* (OLRC, Ken State University). Zugriff am 28.04.2016. Verfügbar unter http://literacy.kent.edu/Oasis/Pubs/0200-14.htm

Reber, K. (2014). Störungen des Schrifspracherwerbs: Lese- und Rechtschreibstörungen. In M. Grohnfeldt (Hrsg.), *Grundwissen der Sprachheilpädagogik und Sprachtherapie* (S. 230-236). Stuttgart: Kohlhammer.

Reber, K. & Schönauer-Schneider, W. (2014). Unterricht und Therapie: sprachheilpädagogischer Unterricht. In M. Grohnfeldt (Hrsg.), *Grundwissen der Sprachheilpädagogik und Sprachtherapie* (S. 323-330). Stuttgart: Kohlhammer.

Rentzsch, K. & Schütz, A. (2009). *Psychologische Diagnostik: Grundlagen und Anwendungsperspektiven*. Stuttgart: Kohlhammer.

Resing, W. C., Stevenson, C. E. & Bosma, T. (2012). Dynamic testing: Measuring inductive reasoning in children with developmental disabilities and mild cognitive impairments. *Journal of Cognitive Education and Psychology, 11* (2), 159-178.

Resing, W. C., Tunteler, E., de Jong, F. M. & Bosma, T. (2009). Dynamic testing in indigenous and ethnic minority children. *Learning and Individual Differences, 19* (4), 445-450.

Resing, W. C., Xenidou-Dervou, I., Steijn, W. M. & Elliott, J. G. (2012). A "picture" of children's potential for learning: Looking into strategy changes and working memory by dynamic testing. *Learning and Individual Differences, 22* (1), 144-150.

Retelsdorf, J. & Möller, J. (2008). Entwicklungen von Lesekompetenz und Lesemotivation Schereneffekte in der Sekundarstufe? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 40* (4), 179-188.

Richardson, M., Abraham, C. & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138* (2), 353.

Richter, T. & Christmann, U. (2002). Lesekompetenz: Prozessebenen und interindividuelle Unterschiede. In N. Groeben & B. Hurrelmann (Hrsg.), *Lesekompetenz. Bedingungen, Dimensionen, Funktionen* (S. 25-58). Weinheim und München: Juventa Verlag.

Rinck, M. (2000). Situationsmodelle und das Verstehen von Erzähltexten: Befunde und Probleme. *Psychologische Rundschau, 51* (3), 115-122.

Rinck, M. & Weber, U. (2003). Who when where: An experimental test of the event-indexing model. *Memory & Cognition, 31* (8), 1284-1292.

Rinck, M., Williams, P., Bower, G. H. & Becker, E. S. (1996). Spatial situation models and narrative understanding: Some generalizations and extensions. *Discourse Processes, 21* (1), 23-55.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? *Psychologische Rundschau, 57* (2), 69-86.

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: the homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality, 21* (5), 667-706.

Rosebrock, C. & Nix, D. (2013). *Grundlagen der Lesedidaktik und der systematischen schulischen Leseförderung* (6. Aufl.). Baltmannsweiler: Schneider Hohengehren.

Rost, D. H. (1987). Leseverständnis oder Leseverständnisse? *Zeitschrift für Pädagogische Psychologie, 1* (3), 175-196.

Rost, D. H. & Schilling, S. R. (2006). Leseverständnis. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (Schlüsselbegriffe, 3., überarb. und erw. Aufl, S. 450-460). Weinheim [u.a.]: Beltz, PVU.

Rost, D. H. & Sparfeldt, J. R. (2007). Leseverständnis ohne Lesen? *Zeitschrift für Pädagogische Psychologie, 21* (3/4), 305-314.

Rubin, H. J. & Rubin, I. S. (2011). *Qualitative interviewing: The art of hearing data*. Los Angeles [u.a.]: Sage.

Rudolph, U. (2003). *Motivationspsychologie*. Weinheim [u.a.]: Beltz, PVU.

Rutland, A. & Campbell, R. (1995). The validity of dynamic assessment methods for children with learning difficulties and nondisabled children. *Journal of Cognitive Education, 5* (1), 81-94.

Samuels, M., Tzuriel, D. & Malloy-Miller, T. (1989). Dynamic assessment of children with learning difficulties. In R. I. Brown & M. Chazan (Hrsg.), *Learning difficulties and emotional problems* (S. 145-166). Calgary: Detselig Enterprises.

Sarges, A. (2013). Lernpotenzial. In M. A. Wirtz & J. Strohmer (Hrsg.), *Dorsch. Lexikon der Psychologie* (16. vollst. überarb. Aufl., S. 951-952). Bern: H. Huber.

Schabmann, A., Landerl, K., Bruneforth, M. & Schmidt, B. M. (2012). Lesekompetenz, Leseunterricht und Leseförderung im österreichischen Schulsystem. *Nationaler Bildungsbericht Österreich, 2,* 17-69.

Schaffner, E. (2009). Determinanten des Leseverstehens. In W. Lenhard & W. Schneider (Hrsg.), *Diagnostik und Förderung des Leseverständnisses* (Tests und Trends, Bd. 7, S. 19-44). Göttingen [u.a.]: Hogrefe.

Schermelleh-Engel, K. & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 119-142). Berlin [u.a.]: Springer.

Schmalhofer, F., McDaniel, M. A. & Keefe, D. (2002). A unified model for predictive and bridging inferences. *Discourse Processes, 33* (2), 105-132.

Schmalhofer, F. & Glavanov, D. (1986). Three components of understanding a programmer's manual: verbatim, propositional, and situational representations. *Journal of memory and language, 25,* 279-294.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik*. Berlin: Springer.

Schmidthals, K. (2005). *Wie das Selbst unser Wissen formt. Der Einfluss independenten und interdependenten Selbstwissens auf die Kontextabhängigkeit mentaler Repräsentationen*. Dissertation. Freie Universität Berlin.

Schmitt, N., Jiang, X. & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95* (1), 26-43.

Schnotz, W. (2006). Was geschieht im Kopf des Lesers? Mentale Konstruktionsprozesse beim Textverstehen aus der Sicht der Psychologie und der kognitiven Linguistik. In H. Blühdorn, E. Breindl & U. H. Waßner (Hrsg.), *Text-Verstehen. Grammatik und darüber hinaus* (S. 222-238). Berlin: De Gruyter.

Schroeder, N. L., Adesope, O. O. & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research, 49* (1), 1-39.

Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs. (2016). *Basic structure of the education system in the Federal Republic of Germany*. Bonn.

Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany. (2015). *The education system in the Federal Republic of Germany 2013/2014: A description of the responsibilities, structures and developments in education policy for the exchange of information in Europe*. Bonn. Verfügbar unter https://www.kmk.org/fileadmin/Dateien/pdf/Eurydice/Bildungswesen-engl-pdfs/secondary.pdf

Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N. & Barterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools, 50* (5), 489-499.

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4* (1), 27-41.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Beschlüsse der Kultusministerkonferenz: Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4)*. München.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2015). *Das Bildungswesen in der Bundesrepublik Deutschland 2013/2014: Darstellung der Kompetenzen, Strukturen und bildungspolitischen Entwicklungen für den Informationsaustausch in Europa*. Bonn.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J. R., Bergmann, P., Birkner, K. et al. (2009). Gesprächsanalytisches Transkriptionssystem 2

(GAT 2). *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion, 10,* 353-402.

Seuring, V. A. (2010). *Förderung des Leseverständnisses mit Methoden des reziproken Lehrens: Effekte unterrichtsintegrierter Trainings für Schülerinnen und Schüler der 5. Klasse*. Dissertation. Justus-Liebig Universität Gießen.

Shafer, K. & Lohse, B. (2005). How to conduct a cognitive interview: A nutrition education example. *US Department of Agriculture, National Institute of Food and Agriculture*.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78* (1), 153-189.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A. & Davis, S. (2009). Effective beginning reading programs: A best-evidence synthesis. *Review of Educational Research, 79,* 1391-1466.

Slavin, R. E., Cheung, A., Groff, C. & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly, 43* (3), 290-322.

Sliwka, A. (2010). From homogeneity to diversity in German education. *Education Teachers for Diversity-meeting the Challenge*.

Snow, C. E. & Verhoeven, L. T. (2001). *Literacy and motivation. Reading engagement in individuals and groups*. Mahwah, N.J: L. Erlbaum.

Souvignier, E., Trenk-Hinterberger, I., Adam-Schwebe, S. & Gold, A. (2008). *Frankfurter Leseverständnistest für 5. und 6. Klassen. (FLVT)*. Göttingen [u.a.]: Hogrefe.

Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology, 84,* 353-363.

Speece, D. L., Cooper, D. H. & Kibler, J. M. (1990). Dynamic assessment, individual differences, and academic achievement. *Learning and Individual Differences, 2* (1), 113-127.

Spielberger, C. D. & Vagg, P. R. (1995). *Test anxiety: Theory, assessment, and treatment:* Taylor & Francis.

Spika, V. (2015). *Die Kybernetische Methode im Schriftspracherwerb: Eine Wirksamkeitsstudie unter besonderer Berücksichtigung von Kindern mit ungünstiger Lernausgangslage*. Dissertation. Universität Augsburg.

SRH Stephen-Hawking-Schule Neckargemünd. (2016). *Grundschule – SRH Stephen-Hawking-Schule*. Zugriff am 22.01.2016. Verfügbar unter http://www.stephenhawkingschule.de/de/unser-angebot/grundschule/

Stanfa, K. M. (2010). *Differentiating among students: the value added of a dynamic assessment of morphological problem-solving*. Dissertation. University of Pittsburgh.

Stanovich, K. E. & Cunningham, A. E. (1991). Reading as constrained reasoning. In R. J. Sternberg & P. A. Frensch (Hrsg.), *Complex problem solving: Principles and mechanisms* (S. 3-60). Hilldale, NJ: Lawrence Erlbaum Associates. Zugriff am 19.09.2013.

Statistisches Bundesamt. (2014a). *Bevölkerung mit Migrationshintergrund - Ergebnisse des Mikrozensus 2013*. Wiesbaden.

Statistisches Bundesamt. (2014b). *Schulen auf einen Blick, Ausgabe 2014*. Wiesbaden.

Stein, H. (1993). Zur Entwicklung eines allgemein-psychologisch begründeten adaptiven Lerntests mit verbalen Analogien (ADANA). In H.-P. Langenfeld & H.-P. Trolldenier (Hrsg.), *Pädagogisch-Psychologische Diagnostik* (S. 125-144). Heidelberg: Asanger.

Steinmayr, R., Crede, J., McElvany, N. & Wirthwein, L. (2015). Subjective well-being, test anxiety, academic achievement: testing for reciprocal effects. *Frontiers in psychology*, *6*, 1994.

Stern, E. (2001). Intelligence, prior knowledge, and learning. *International encyclopedia of the social and behavioral sciences, 11,* 7670-7674.

Stern, E. & Neubauer, A. (2016). Intelligenz: kein Mythos, sondern Realität. *Psychologische Rundschau, 67* (1), 15-27.

Sternberg, R. J. (2000). The concept of intelligence. In R. J. Sternberg (Hrsg.), *Handbook of intelligence* (S. 3-15). Cambridge University Press.

Sternberg, R. J. (2004). North American approaches to intelligence. In R. J. Sternberg (Hrsg.), *International handbook of intelligence* (S. 411-444). Cambridge, UK, New York, NY [u.a.]: Cambridge University Press.

Sternberg, R. J. & Grigorenko, E. L. (2002). *Dynamic testing. The nature and measurement of learning potential*. Cambridge et al.: Cambridge University Press.

Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C. et al. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence, 30* (2), 141-162.

Südkamp, A., Kaiser, J. & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743-762.

Suh, S. & Trabasso, T. (1993). Inferences during reading: Converging evidence from discourse analysis, talk-aloud protocols, and recognition priming. *Journal of memory and language, 32,* 279-300.

Süß, H.-M. (2005). Experimentelle Methoden. In H. Weber (Hrsg.), *Handbuch der Persönlichkeitspsychologie und differentiellen Psychologie* (Handbuch der Psychologie, Bd. 2, S. 166-180). Göttingen: Hogrefe.

Swanson, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology, 84* (4), 473-488.

Swanson, H. L. (1995). Effects of dynamic testing on the classification of learning disabilities: The predictive and discriminant validity of the Swanson-Cognitive Processing Test (S-CPT). *Journal of Psychoeducational Assessment, 13* (3), 204-229.

Swanson, H. L. (2010). Does the dynamic testing of working memory predict growth in nonword fluency and vocabulary in children with reading disabilities? *Journal of Cognitive Education and Psychology, 9* (2), 139-165.

Swanson, H. L. (2011). Dynamic testing, working memory, and reading comprehension growth in children with reading disabilities. *Journal of Learning Disabilities, 44* (4), 358-371.

Swanson, H. L. & Howard, C. B. (2005). Children with reading disabilities: Does dynamic assessment help in the classification? *Reading Disability Quarterly, 28* (1), 17-34.

Taboada, A., Tonks, S. M., Wigfield, A. & Guthrie, J. T. (2009). Effects of motivational and cognitive variables on reading comprehension. *Reading and Writing, 22* (1), 85-106.

Tarchi, C. (2010). Reading comprehension of informative texts in secondary school: A focus on direct and indirect effects of reader's prior knowledge. *Learning and Individual Differences, 20* (5), 415-420.

Tent, L. (2006). Zensuren. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (Schlüsselbegriffe, 3., überarb. und erw. Aufl, S. 873-879). Weinheim [u.a.]: Beltz, PVU.

Testangst. (2013). In M. A. Wirtz & J. Strohmer (Hrsg.), *Dorsch. Lexikon der Psychologie* (16. vollst. überarb. Aufl., S. 1539). Bern: H. Huber.

Therriault, D. J. & Rinck, M. (2007). Multidimensional situation model. In F. Schmalhofer & C. A. Perfetti (Hrsg.), *Higher Level Language Processes in the Brain* (S. 311-327). Mahwah, NJ: Erlbaum.

Therriault, D. J., Rinck, M. & Zwaan, R. A. (2006). Assessing the influence of dimensional focus during situation model construction. *Memory & Cognition, 34* (1), 78-89.

Tischler, T., Daseking, M. & Petermann, F. (2013). Effekt der Schulform auf die Entwicklung der Lesegeschwindigkeit. *Diagnostica, 59* (4), 215-226.

Toutenbourg, H. (1994). *Versuchsplanung und Modellwahl*. Berlin [u.a.]: Springer-Verlag.

Tsai, Y.-C. & Li, Y.-C. (2012). Test anxiety and foreign language reading anxiety in a reading-proficiency test. *Journal of Social Sciences, 8* (1), 95-103.

Tzuriel, D. & Klein, P. S. (1985). The assessment of analogical thinking modifiability among regular, special education, disadvantaged, and mentaly retarded children. *Journal of Abnormal Child Psychology, 13* (4), 539-552.

Tzuriel, D. & Shamir, A. (2002). The effects of mediation in computer assisted dynamic assessment. *Journal of Computer Assisted Learning, 18,* 21-32.

Van den Broek, P. (2010). Using texts in science education: Cognitive processes and knowledge representation. *Science, 328* (5977), 453-456.

Van den Broek, P. & Gustafson, M. (1999). Comprehension and memory for texts: Three generations of reading research. In S. R. Goldman, A. C. Graesser & P. van den Broek (Hrsg.), *Narrative comprehension, causality, and coherence. Essays in honor of Tom Trabasso* (S. 15-34). Mahwah, N.J: L. Erlbaum Associates.

Van den Broek, P. & Lorch, R. F. (1993). Network representations of causal relations in memory for narrative texts: Evidence from primed recognition. *Discourse Processes, 16* (1-2), 75-98.

Van der Veer, R. & Valsiner, J. (1994). *The Vygotsky reader*. Oxford, UK: Blackwell.

Van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York [u.a.]: Acad. Pr.

Van Someren, M. W., Barnard, Y. F. & Sandberg, J. A. C. (1994). *The think aloud method. A practical guide to modelling cognitive processes*. London: Academic Press.

Veletsianos, G. & Russell, G. S. (2013). What do learners and pedagogical agents discuss when given opportunities for open-ended dialogue? *Journal of Educational Computing Research, 48* (3), 381-401.

Völzke, K. (2012). *Lautes Denken bei kompetenzorientierten Diagnoseaufgaben zur naturwissenschaftlichen Erkenntnisgewinnung*. (Vol. 20). Kassel: kassel university press.

Von Goldammer, A. (2010). *Von der Sprache zur Schriftsprache. Diagnostische und prognostische Validität der Erfassung von*

Vorläuferkompetenzen der Schriftsprache im Vorschulalter. Dissertation. Universität Hildesheim.

Weidner, J. (2014). *Diagnostik und Förderung von Leseverständnis*. Wissenschaftliche Hausarbeit. Pädagogische Hochschule Heidelberg.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in der Schule – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessung in der Schule* (S. 17-31). Weinheim, Basel: Beltz.

Wiberg, M. & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation, 14* (5), 2.

Wieczerkowski, W., Nickel, H., Janowski, A., Fittkau, B. & Rauer W. (1980). *Angstfragebogen für Schüler (AFS)*. Göttingen: Westermann.

Wiedl, K. H. (1984). *Lerntests: nur ein Forschungsgegenstand?* Psychologische Forschungsberichte aus dem Fachbereich 8 der Universität Osnabrück. Nr. 35. Universität Osnabrück.

Wiedl, K. H. & Carlson, J. S. (1981). Dynamisches Testen bei lernbehinderten Sonderschülern mit dem farbigen Matrizentest von Raven. *Heilpädagogische Forschung, 11,* 19-26.

Willenberg, H. (2007). Wortschatz. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Konsequenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (S. 130-139). Weinheim [u.a.]: Beltz.

Willis, G. B. (1999). *Cognitive interviewing. A "how to" guide,* Research Triangle Institute. Verfügbar unter http://appliedresearch.cancer.gov/archive/cognitive/interview.pdf

Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76* (2), 92-104.

Wolschke, P., Wilmes, K., Huber, W. & Guthke, J. (1995). Unterschiedliches Lernverhalten von Kindern im ersten Schuljahr beim begriffsanalogen

Klassifizieren im Leipziger Lerntest DP-BAK. *Heilpädagogische Forschung, 21,* 97-110.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ConQuest (Version 2.0) [Computer software and manual].

Wygotsky, L. (1978). Interaction between learning and development. *Readings on the development of children, 23* (3), 34-41.

Zwaan, R. A. (1996). Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22* (5), 1196-1207.

Zwaan, R. A., Magliano, J. P. & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 386-397.

Zwaan, R. A., Radvansky, G. A., Hilliard, A. E. & Curiel, J. M. (1998). Constructing multidimensional situation models during reading. *Scientific Studies of Reading,* 199-220.

Zwaan, R. A., Langston, M. C. & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6,* 292-297.

Zwaan, R. A. & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123* (2), 162-185.

Zwaan, R. A. & van Oostendorp, H. (1993). Do readers construct spatial representations in naturalistic story comprehension? *Discourse Processes, 16* (1-2), 125-143.